Theoretical and Applied Genetics (2022) 135:2891–2905
https://doi.org/10.1007/s00122-022-04157-1

ORIGINAL ARTICLE

Assessing the response to genomic selection by simulation

Harimurti Buntaran[1] · Angela Maria Bernal-Vasquez[2] · Andres Gordillo[3] · Morten Sahr[3] · Valentin Wimmer[2] · Hans-Peter Piepho[1]

Received: 24 January 2022 / Accepted: 16 June 2022 / Published online: 14 July 2022

Abstract

Key message **We propose a simulation approach to compute response to genomic selection on a multi-environment framework to provide breeders the number of entries that need to be selected from the population to have a defined probability of selecting the truly best entry from the population and the probability of obtaining the truly best entries when some top-ranked entries are selected.**

Abstract The goal of any plant breeding program is to maximize genetic gain for traits of interest. In classical quantitative genetics, the genetic gain can be obtained from what is known as "Breeder's equation". In the past, only phenotypic data were used to compute the genetic gain. The advent of genomic prediction (GP) has opened the door to the utilization of dense markers for estimating genomic breeding values or GBV. The salient feature of GP is the possibility to carry out genomic selection with the assistance of the kinship matrix, hence improving the prediction accuracy and accelerating the breeding cycle. However, estimates of GBV as such do not provide the full information on the number of entries to be selected as in the classical response to selection. In this paper, we use simulation, based on a fitted mixed model for GP in a multi-environmental framework, to answer two typical questions of a plant breeder: (1) How many entries need to be selected to have a defined probability of selecting the truly best entry from the population; (2) what is the probability of obtaining the truly best entries when some top-ranked entries are selected.

Introduction

In plant breeding programs, the breeder's equation (Lush 1942) has been central to measure response to selection, known as genetic gain. The genetic gain also describes the breeding value of a population in one cycle of selection for the trait of interest (Rutkoski 2019). Response to selection is based on the heritability and the selection differential. Hence, a trait with a high heritability and a high selection differential can be considered to have a large genetic gain. The selection differential *per-se* is based on the number of selected entries, which mainly relies on the breeder's eyes when the selection is solely based on the phenotype. Environmental effects and genotype × environment interactions play a role in masking the true genotype value. This is why a breeding program is never conducted in a single environment (Crossa et al. 2017). The imbalance prevailing in datasets from the multi-environmental trials (MET) brings additional complexity to the data analysis for estimating response to selection.

The advent of genomic prediction (GP) allows breeders to estimate genomic-based breeding values via dense markers (Meuwissen et al. 2001). The salient feature of the GP is the improvement in the accuracy of breeding value estimation by exploiting the kinship matrix. Lorenz et al. (2011) and Crossa et al. (2017) provided a comprehensive review for the implementation and benefit of GP as the basis of genomic selection in plant breeding. Although GP allows for selection to happen earlier and provides better accuracy in breeding value estimation, the incorporation of GEI to the GP framework is still a challenge.

Genomic prediction is by now routinely used as the basis of genomic selection in many plant breeding programs

Communicated by Ian Mackay.

✉ Hans-Peter Piepho
hans-peter.piepho@uni-hohenheim.de

[1] Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstraße 23, 70599 Stuttgart, Germany

[2] KWS SAAT SE Co. KGaA, Grimsehlstaße. 31, 37574 Einbeck, Germany

[3] KWS-LOCHOW GmbH, Ferdinand-von-Lochow-Straße 5, 29303 Bergen, Germany

worldwide. The usual approach for evaluating the predictive accuracy of GP methods is to compute the correlation between observed and predicted genomic breeding values (GBV) using cross-validation. While this method is very useful in comparing alternative methods and designs, it does not usually give the plant breeder the full picture needed to decide on the number of breeding entries to be selected. Typical questions a plant breeder has in this regard may be exemplified as follows: (i) How many entries do I need to select to have a defined certainty (probability) of selecting the truly best entry from my population? (ii) If I select the n top-ranked entries, what is the probability of picking the $m \leq n$ truly best entries from the population?

Predictive accuracy, useful as it undoubtedly is, does not give direct answers to such crucial questions. These probabilities are hard to compute analytically for various reasons, including the imbalance of the data and the complexities of the mixed model used for analysis. The easiest and also the most tangible way to compute them is by simulation based on the fitted model. This idea was outlined and illustrated by examples in Piepho and Möhring (2007), and it was also used, though in slightly different context, in Piepho and van Eeuwijk (2002) and Kleinknecht et al. (2016). However, neither of these applications involved the use of marker data for GP. Here, we illustrate the use of this method for GP in a MET framework using an example from a hybrid rye breeding program.

Materials and methods

A rye example

Description of the dataset and underlying population structure

The phenotypic data are proprietary of a commercial hybrid rye breeding program by KWS LOCHOW established in central Europe. In the program, the seed and pollen gene pools are developed and tested independently. After crossing, a few generations of single plant selfing, selection and per se performance line selection, testcrosses are produced between inbred lines and two testers from the opposite heterotic pool.

The testcrosses are submitted to field trials. In the first year of general combining ability (GCA) evaluation (hereafter GCA1 trials), entries (i.e. testcrosses) are planted in several locations. A subset of entries is selected and forwarded to a second year of field evaluations (hereafter GCA2 trials). The selected fraction from the GCA2 trials is evaluated on the field in a third year (hereafter GCA3 trials). The entries of the GCA3 trials are the testcrosses developed in the previous year. A sequence of GCA1 to GCA3 trials constitutes a selection cycle. All trials within a location are laid out as α-designs with two replicates. Alpha (α) designs are a class of generalized lattice designs generated based on alpha arrays (Patterson and Williams 1976; Williams et al. 2002). The trial network follows a sparse pattern, where subsets of entries are evaluated in series of trials in a given subset of locations but trying to cover as many locations as possible. Sparse testing is described in Jarquín et al. (2014).

Figure 1 shows the selection cycle structure of the rye hybrid breeding program. In each cycle, there are 3 years of tests for GCA. A selection is conducted each year. Thus, in each cycle, the number of entries decreases from GCA1 to GCA3.

In this study, two breeding pools were used, i.e. seed pool and pollen pool. The data available for both pools cover the years 2016 to 2020. Thus, there were three complete selection cycles available, i.e. Cycle 1 (2016–2018), Cycle 2 (2017–2019), and Cycle 3 (2018–2020). For Cycle 4, the available dataset comprised only GCA1 (2019) and GCA2 (2020), and for Cycle 5, only GCA1 (2020) was available.

Marker data

All genotypes from the seed and pollen pools were genotyped with an Illumina INFINIUM chip with 9963 single-nucleotide polymorphisms (SNPs) (KWS SAAT SE & Co. KG, Einbeck, Germany). The SNP used across the years partially overlap the 600 k-SNP assay of Bauer et al. (2017) and the 5 k-SNP assay of Martis et al. (2013). Monomorphic markers and markers with a minor allele frequency (MAF) < 0.55, or > 10% missing values per marker were dropped. The marker cleaning was done using ASRgenomics package version 1.0.0 (Gezan et al. 2021) implemented in R (R Core Team 2021). The final number of SNP markers used for our analyses for the seed pool was 6246 and 7716 for the pollen pool.

Population and statistical models

GCA1 response to selection assessment

The routine analysis we envision here is based on the current year's dataset only, which in this study was taken to be the year 2020. This type of analysis is commonly done in many breeding programs, where the time between data acquisition from the current trials and selection decisions is limited. This approach also reflects the fact that selection decisions are made each year only for the entries tested in that year. We are focusing on GCA1 here, for which there are no data on the same entries from previous years.

A key challenge is that a standard single-year analysis cannot dissect the GBV × year interaction effects from GBV main effects. In particular, response to selection

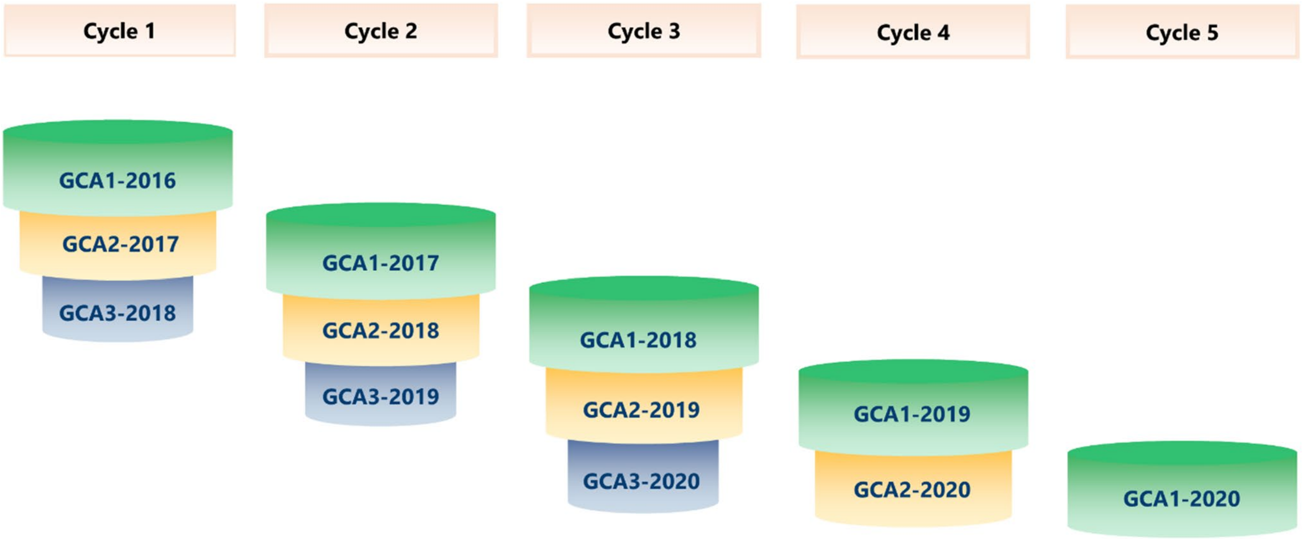

Fig. 1 The structure of selection cycles in the rye hybrid breeding program. The number of entries decreases due to selection in each GCA trial. In each cycle, inbred lines are crossed with two testers of the opposite gene pool

based on such an analysis may be over-estimated because it is based on the sum of these two effects, whereas only the first component contributes to selection response in relation to future performance of the selected entries. However, if data from multiple years are available, variance components for these two effects can be estimated, and a key idea of this paper is to use such estimates for single-year analysis to dissect GBV and GBV × year effects.

Estimation of the GBV × year variance from multi-year data can be done a priori and then plugged into the analysis for the data from the current year. In principle, we could also plug in the long-term estimate for the GBV variance estimate, but we prefer to use the current GBV variance to adjust for population structure in the current year. The major effect of the inclusion of a GBV × year interaction effect in a single-year analysis is an increased shrinkage of the genomic best linear unbiased predictions (GBLUPs) of the GBV main effects.

Moreover, the simulation based on this model will yield a smaller simulated response to selection. It is stressed here that this simulated response to selection is more realistic, as it properly reflects the fact that the apparent GBV main effect seen in a specific year based on the usual method of analysis is, in fact, the sum of the true GBV main effect and the GBV × year interaction effect for that year. In this study, the current-year (GCA1-2020) analysis uses estimates of the variances for GBV main effects and GBV × year interaction effects, which are obtained from the multi-year analysis based on the previous years. The steps for the multi-year analysis are, therefore, as follows:

1. Estimate $\rho = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{gy}^2}$ from long-term data, where σ_g^2 and σ_{gy}^2 are variances of the GBV (explained in more detail in subsection CYC and MY models) and of the GBV-by-year interaction.
2. Estimate the apparent GBV variance $\sigma_{\tilde{g}}^2$ of the current year (GCA1-2020), using a model that only has a GBV main effect but no GBV × year interaction effect. It may be assumed that $\sigma_{\tilde{g}}^2 = \sigma_g^2 + \sigma_{gy}^2$.
3. Multiply this estimate of $\sigma_{\tilde{g}}^2$ by the estimate of ρ to obtain an estimate of σ_g^2 for the current year, and multiply by an estimate of (1 − ρ) to obtain an estimate of σ_{gy}^2 for the current year.
4. Rerun the model for GCA1-2020 data by plugging in and fixing σ_g^2 and σ_{gy}^2 at their estimates from step 3, and using the other variance estimates obtained previously from the GCA1-2020 model. The error variance–covariance matrix of GBLUPs obtained from the rerun model will be used for the simulation, as described in more detail in subsection The general simulation approach.

A key challenge with this approach is how ρ in the first step above can be estimated in the multi-year analysis. Here, we propose two alternative analyses for this purpose, i.e. a combined-cycles (CYC) and a multi-year analysis based on GCA1 data of the years 2016 to 2019 (MY). Figure 2 shows the differences in dataset structure between the CYC and MY analyses. Applying the CYC analysis, we used three complete selection cycles and one incomplete cycle,

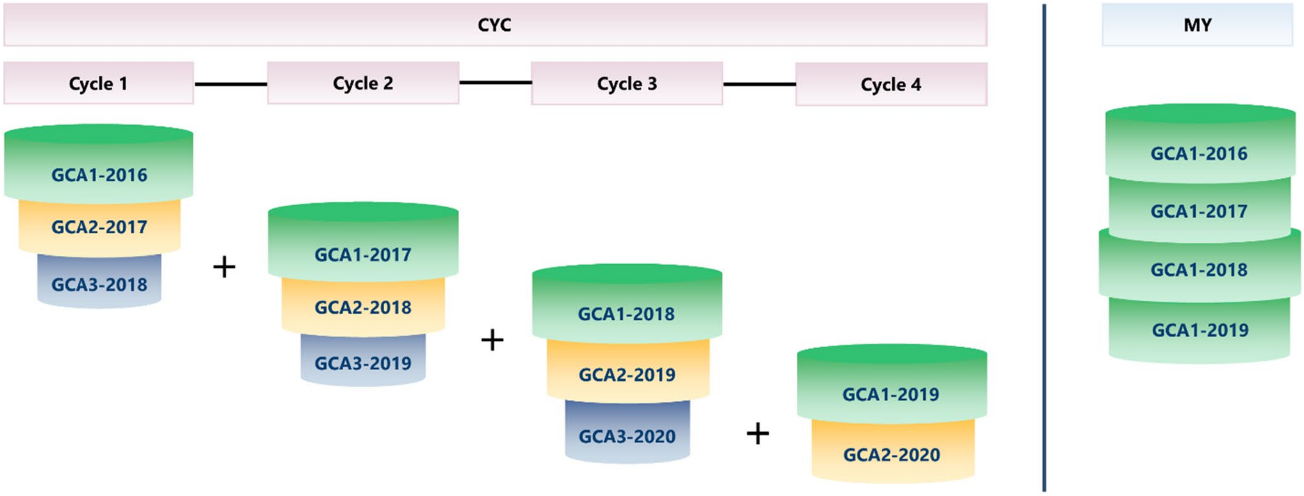

Fig. 2 Illustration of datasets used for the CYC and MY analyses. In CYC, the GCA1, GCA2, and GCA3 datasets from all cycles are combined. In MY, the dataset comprises the GCA1 data of the years 2016 to 2019

Table 1 The comparisons between CYC and MY analyses in terms of analysis strategy and datasets

Properties	CYC	MY
Analysis strategy	Obtain a single ρ estimate from combined Cycle 1 to Cycle 4 data for the estimates of σ_g^2 and σ_{gy}^2 adjustment in the GCA1 (2020) analysis	Obtain a single ρ estimate from GCA1 data of the years 2016 to 2019 for the estimates of σ_g^2 and σ_{gy}^2 adjustment in the GCA1 (2020) analysis
Datasets connectivity	The datasets have a connectivity within a cycle	Using only the GCA1 dataset for all four years

as shown in Fig. 2. All entries in each GCA trial were used in the CYC analysis. On the other hand, the MY analysis only used the GCA1 datasets from the years 2016 to 2019. The checks were removed in both datasets. Table 1 summarizes the comparisons between the CYC analysis and the MY analyses regarding data handling, analysis strategy, and dataset. Moreover, total entry number from all cycles per pool is given in Table 2.

CYC and MY models

For the CYC and MY analyses, a two-stage approach was used. In the CYC analysis, the approach was applied for each cycle, while in the MY analysis, the approach was applied directly in the four-year dataset. In Stage 1, the entry means per year across locations were computed. Thus, the following linear mixed model was fitted per year for phenotypic analysis at the plot level:

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{X_g}\boldsymbol{\beta}_\mathbf{g} + \mathbf{Z_l u_l} + \mathbf{Z_t u_t} + \mathbf{Z_{gt} u_{gt}} + \mathbf{Z_{gl} u_{gl}} + \mathbf{Z_{ls} u_{ls}} + \mathbf{Z_{lsr} u_{lsr}} + \mathbf{Z_{lsrb} u_{lsrb}} + \mathbf{e} \quad (1)$$

where **Y** is the vector of observed plot yields, **1** is a vector of ones, μ is the general mean, **X** is the design matrix relating to fixed effects of the entry (g), **Z** is the incidence matrix relating to random effects followed with the subscripts t, l, s, r, and b relating to the factors tester, location, trial within location, replicate within trial, and block within replicate, respectively, and **e** is the residual associated with the observation **Y**. The distributional assumption for each random effect, $\mathbf{u_l}, \mathbf{u_t}, \mathbf{u_{gt}}, \mathbf{u_{gl}}, \mathbf{u_{ls}}, \mathbf{u_{lsr}}$, and $\mathbf{u_{lsrb}}$, was a Gaussian distribution with zero mean, independence of individual effects, and constant variance, as delineated in Appendix (Table 12). The entry effect, $\boldsymbol{\beta}_\mathbf{g}$, was fixed in Stage 1 to obtain the adjusted entry means via generalized least squares, and so avoid double shrinkage (Smith et al. 2001; Piepho et al. 2012). Thus, the adjusted entry means are empirical best linear unbiased estimates (EBLUEs) of the entries' expected values under the assumed model. The residual variance structure was heterogeneous with location-specific variance and independent error effects, as described in Appendix (Table 12).

Table 2 The total entry number from all cycles per pool

	Total entry number	
	Seed pool	Pollen pool
GCA1	3931	7474
GCA2[a]	353	669
GCA3[a]	27	53

[a] After selection from the previous GCA

In Stage 2, the adjusted entry means were assembled across several years, i.e. four years for the MY analysis and five years for the CYC analysis since it included the GCA2-2020 dataset. The following GP model was fitted at this stage:

$$\mathbf{Y}_{\mathbf{adj}} = \mathbf{1}\upmu + \mathbf{Z}_{\mathbf{y}}\mathbf{u}_{\mathbf{y}} + \mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\mathbf{g}} + \mathbf{Z}_{\mathbf{gy}}\mathbf{u}_{\mathbf{gy}} + \mathbf{e} \tag{2}$$

where $\mathbf{Y}_{\mathbf{adj}}$ is the vector of entry-year means, $\mathbf{Z}_{\mathbf{y}}$ is the incidence matrix of year main effects, $\mathbf{Z}_{\mathbf{g}}$ represents the incidence matrix of entries, $\mathbf{u}_{\mathbf{g}}$ is the vector of GBV, which is defined as $\mathbf{u}_{\mathbf{g}} = \mathbf{Qv}$, where $\mathbf{Q}$ is the $N \times P$ marker genotypes matrix for N entries and P markers, and $\mathbf{v}$ is the vector of marker effects, $\mathbf{v} \sim N(0, \mathbf{I}\sigma^2_{\mathrm{g}})$ where σ^2_g is the genomic variance. Following these assumptions, then we have $\mathbf{u}_{\mathbf{g}} \sim MVN(0, \mathbf{K}\sigma^2_{\mathrm{g}})$ where $\mathbf{K} = \mathbf{QQ}^{\mathbf{T}}$. Here, $\mathbf{Q}$ was mean-centred and scaled according to the VanRaden (2008) method. There are alternative methods for computing genomic relationship matrices such as Astle and Balding (2009), Endelman and Jannink (2012), Yang et al. (2010), and the recent method using average semivariance by Feldmann et al. (2020), but these are all equivalent in terms of the resulting BLUPs of $\mathbf{u}_{\mathbf{g}}$. The key component of Eq. 2 is a GBV × year interaction effect, $\mathbf{u}_{\mathbf{gy}}$. Thus, $\mathbf{Z}_{\mathbf{gy}}$ is a block-diagonal matrix with blocks given by the coefficient of entries in a given year, $\oplus_{j=1}^{J}\mathbf{Z}_{\mathrm{gy}_j}$, where j is a subscript for years, and $\mathbf{u}_{\mathbf{gy}} \sim MVN(0, \mathbf{G}_{\mathrm{gy}})$ is the vector of GBV-by-year effects, where $\mathbf{G}_{\mathrm{gy}} = \oplus_{j=1}^{J}\mathbf{K}_j\sigma^2_{\mathrm{gy}}$, $\mathbf{K}_j$ is the kinship of all entries tested in the jth year, as shown in Appendix (Table 13). The vector $\mathbf{e}$ is the residual term, where $\mathrm{var}(\mathbf{e})$ is approximated using a diagonal variance–covariance matrix as proposed in Smith et al. (2001). The two-stage weighted approach is useful when the single-stage approach burdens the computing time. Furthermore, a cross-validation study by Buntaran et al. (2020) demonstrated that the two-stage approach with Smith's weighting was competitive to the single-stage approach.

Current-year model

A single-stage GP model was used to obtain the apparent (unadjusted) GBV variance ($\sigma^2_{\tilde{\mathrm{g}}}$). The implemented model is Eq. 1 used in Stage 1 of the two-stage approach in the CYC and MY analyses, but the entry effect was replaced with the GBV as follows:

$$\begin{aligned}\mathbf{Y} = \mathbf{1}\upmu + \mathbf{Z}_{\mathbf{l}}\mathbf{u}_{\mathbf{l}} + \mathbf{Z}_{\mathbf{t}}\mathbf{u}_{\mathbf{t}} + \mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\tilde{\mathbf{g}}} + \mathbf{Z}_{\mathbf{gt}}\mathbf{u}_{\mathbf{gt}} \\ + \mathbf{Z}_{\mathbf{gl}}\mathbf{u}_{\mathbf{gl}} + \mathbf{Z}_{\mathbf{ls}}\mathbf{u}_{\mathbf{ls}} + \mathbf{Z}_{\mathbf{lsr}}\mathbf{u}_{\mathbf{lsr}} + \mathbf{Z}_{\mathbf{lsrb}}\mathbf{u}_{\mathbf{lsrb}} + \mathbf{e}\end{aligned} \tag{3}$$

All terms are defined as for Eq. 1 and explained in Appendix (Table 12), while $\mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\tilde{\mathbf{g}}}$ is defined as in Eq. 2 and is explained in Appendix (Table 13). From this model, an estimate of $\sigma^2_{\tilde{\mathrm{g}}}$ was obtained, which in its turn is used to get estimates of σ^2_{g} and σ^2_{gy} by using ρ from the long-term data as described in GCA1 response to selection assessment subsection. Equation 3 was rerun with only one iteration by replacing $\mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\tilde{\mathbf{g}}}$ with $\mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\tilde{\mathbf{g}}} = \mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\mathbf{g}} + \mathbf{Z}_{\mathbf{gy}}\mathbf{u}_{\mathbf{gy}}$. Note that the design matrix for $\mathbf{u}_{\mathbf{gy}}$ is $\mathbf{Z}_{\mathbf{g}}$, which is the basis of the approach suggested here. Thus, the estimate of $\sigma^2_{\tilde{\mathrm{g}}}$ was replaced with the estimates of σ^2_{g} and σ^2_{gy}. Also, in the rerun, the estimates of σ^2_{g} and σ^2_{gy} and other variance estimates of the random effects from Eq. 3 were held fixed at the prespecified values.

GCA2 response to selection assessment

The GCA2 trial is crucial since in this trial the entries will be selected for the final trial of a selection cycle, i.e. the GCA3. Thus, it is desirable to use the GCA1 and GCA2 datasets for the GCA2 analysis. However, there is a challenge in using the previous trials' data, i.e. the GCA1. If all entries from GCA1 are also used, then an estimate of the genetic variance applying to GCA1 is obtained (Piepho and Möhring 2006), whereas a variance estimate applying to GCA2 is needed.

Since the dropped-out entries from GCA1 have no contribution to the genetic variance in GCA2, we selected the common entries that went through to GCA2 from GCA1 as depicted with the transparent red cylinders in Fig. 3. In this assessment, there were four selection cycles available from the main dataset, as shown in Fig. 3. The checks were also removed as in the GCA1 assessment.

GCA2 model

For each cycle, a GP model was implemented using a single-stage approach. This was feasible due to only a relatively small number of entries in the datasets. As in the GCA1 model, the single-stage GP model for the GCA2 has a key component, i.e. a GBV × year interaction effect ($\mathbf{u}_{\mathbf{gy}}$). The single-stage model is fitted in the plot level as follows:

$$\begin{aligned}\mathbf{Y} = \mathbf{1}\upmu + \mathbf{Z}_{\mathbf{y}}\mathbf{u}_{\mathbf{y}} + \mathbf{Z}_{\mathbf{l}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{l}_{(\mathrm{y})}} + \mathbf{Z}_{\mathbf{t}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{t}_{(\mathrm{y})}} \\ + \mathbf{Z}_{\mathbf{g}}\mathbf{u}_{\mathbf{g}} + \mathbf{Z}_{\mathbf{gy}}\mathbf{u}_{\mathbf{gy}} + \mathbf{Z}_{\mathbf{gt}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{gt}_{(\mathrm{y})}} + \mathbf{Z}_{\mathbf{gl}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{gl}_{(\mathrm{y})}} \\ + \mathbf{Z}_{\mathbf{ls}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{ls}_{(\mathrm{y})}} + \mathbf{Z}_{\mathbf{lsr}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{lsr}_{(\mathrm{y})}} + \mathbf{Z}_{\mathbf{lsrb}_{(\mathrm{y})}}\mathbf{u}_{\mathbf{lsrb}_{(\mathrm{y})}} + \mathbf{e}\end{aligned} \tag{4}$$

This single-stage model is used as the year-wise analysis. Thus, the tester (t), location (l), entry × tester (gt), entry × location (gl), trial within location (ls), the replication (lsr), and the incomplete block (lsrb) effects were nested within years (y). Furthermore, the distributional assumption for each random effect, $\mathbf{u}_{\mathbf{y}}$, $\mathbf{u}_{\mathbf{l}_{(\mathrm{y})}}$, $\mathbf{u}_{\mathbf{t}_{(\mathrm{y})}}$, $\mathbf{u}_{\mathbf{gt}_{(\mathrm{y})}}$, $\mathbf{u}_{\mathbf{gl}_{(\mathrm{y})}}$, $\mathbf{u}_{\mathbf{ls}_{(\mathrm{y})}}$, $\mathbf{u}_{\mathbf{lsr}_{(\mathrm{y})}}$,

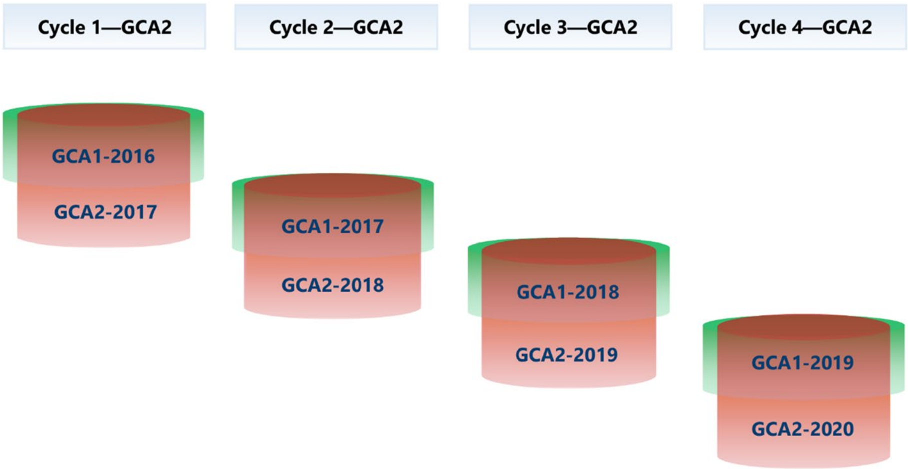

Fig. 3 Illustration of the selected entries for the GCA2 assessment. The transparent red cylinders illustrate the selected common entries in GCA1 and GCA2

and $\mathbf{u}_{\mathbf{lsrb}_{(y)}}$, was Gaussian with zero mean, assuming independence of individual effects and year-specific variance to mimic the year-wise analysis. The GBV main effects $\left(\mathbf{u}_{\mathbf{g}}\right)$ and the GBV × year interaction effects $\left(\mathbf{u}_{\mathbf{gy}}\right)$ had the same variance–covariance structures as described in Eq. 2. For the residual variance, its structure was heterogeneous with year-location-specific variance and independent error effects. The variance–covariance structure for each term in the model is explained in Appendix (Table 14).

The general simulation approach

The fitted linear mixed models have a random vector $\mathbf{g} = \left(\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{\mathrm{N}}\right)^T$ of the GBV, $\mathbf{g}_i$ $(i = 1, \ldots, N)$ of the N entries in the current population from which a selection of a subset of $n < N$ entries is to be performed. The random effects are assumed to be multivariate normal with zero mean and variance–covariance matrix $\mathrm{var}(\mathbf{g}) = \mathbf{K}\sigma_{\mathrm{g}}^2$, where $\mathbf{K}$ is the kinship matrix computed from markers as described in CYC and MY models, and σ_{g}^2 is a genomic variance. The random GBV effect is fitted within a larger linear mixed model accounting for all sources of variation due to the experimental design. In the MET framework, the model for phenotype analysis may be fitted either in a single stage or in several stages, as we demonstrate in this study. In any case, the random effects $\mathbf{g}$ will ultimately be estimated based on the GBLUPs, $\widehat{\mathbf{g}}$, from the fitted linear mixed models.

The key of the simulation approach is to simulate a large number S of realizations of the genetic effects of interest, $\mathbf{g}$, and the corresponding estimated GBLUPs, $\widehat{\mathbf{g}}$, from their joint distribution and determine any quantity of interest related to the response to selection from this simulated distribution. We here use the fact that the joint distribution of $\mathbf{g}$ and $\widehat{\mathbf{g}}$ is multivariate normal with zero mean and variance–covariance matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{D} & \mathbf{M} \\ \mathbf{M} & \mathbf{M} \end{bmatrix} \tag{5}$$

where $\mathbf{M} = \mathrm{var}\left(\widehat{\mathbf{g}}\right)$ is the unconditional variance of $\widehat{\mathbf{g}}$, $\mathbf{M} = \mathbf{D} - \mathbf{C}$, $\mathbf{D} = \mathbf{K}\sigma_{\mathrm{g}}^2$, and the $\mathbf{C} = \mathrm{var}\left(\widehat{\mathbf{g}} - \mathbf{g}\right)$, which can be obtained routinely from the inverse of the coefficient matrix of the mixed model equations (MME) (McLean et al. 1991; Piepho and Möhring, 2007).

We use a decomposition of the Ω matrix given by:

$$\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' \tag{6}$$

This decomposition can be obtained by Cholesky decomposition, in case the matrix Ω is positive-definite, or using a singular value decomposition (SVD) in the case of the Ω is not positive-definite.

We then simulate the values of $\mathbf{w}' = (\mathbf{g}', \widehat{\mathbf{g}}')$ by:

$$\mathbf{w} = \boldsymbol{\Gamma}\mathbf{z} \tag{7}$$

where $\mathbf{z}$ is a $2N$-random vector drawn from a standard normal distribution, where N is the size of $\mathbf{g}$. For a single simulation of $\mathbf{w}$, the vector $\mathbf{z}$ is generated using a random number generator, e.g. based on the Box-Muller method with the `rannor()` function in SAS, or the `rnorm()` function in R. The simulations were conducted in R 4.1.0 (R Core Team 2021) after the GCA1 and GCA2 models were fitted in ASReml-R 4.1.0.160 (Butler et al. 2017) using RStudio (RStudio Team 2021). The R codes for fitting the

GCA1 models are provided in the electronic supplementary materials.

Quantities of interest estimated from the simulated distribution

For each of the S simulation runs, the n best entries based on their GBLUPs $\hat{\mathbf{g}}$ were selected. Then, across all S simulation runs, we determined the proportion of cases where the selected set of n entries containing the m truly best ones based on the associated true genetic values in $\mathbf{g}$. This was done for a range of values for n and $m \leq n$. The probability plots showing the proportion of cases in which the selected set of n entries contain the m truly best one were generated using the `ggplot2` package (Wickham, 2016). Additionally, Pearson's product-moment correlations between the $\hat{\mathbf{g}}$ and $\mathbf{g}$, and between $rank(\mathbf{g})$ and $rank\left(\hat{\mathbf{g}}\right)$ were computed. The benefit of correlations based on the ranks is that they are not affected by the extreme values of $\hat{\mathbf{g}}$ and $\mathbf{g}$. The R code for conducting the simulation and generating the probability plot is provided in the electronic supplementary materials.

Results

GCA1 response to selection assessment

In the GCA1 assessment, the performance of the current-year model depends on the value of $\hat{\rho}$ from the multi-year analysis. The estimate $\hat{\rho}$ depends on the relative size of the variance estimates of GBV $\left(\sigma^2_g\right)$ and GBV × year interaction effects $\left(\sigma^2_{gy}\right)$. The variance estimates for year main effects $\left(\sigma^2_y\right)$, GBV main effects $\left(\sigma^2_g\right)$, the GBV × year interaction effects $\left(\sigma^2_{gy}\right)$, and the estimates of $\hat{\rho}$ along with their associated standard errors in the CYC and MY analyses based on Eq. 2 are summarized in Table 3. As expected, the estimates of σ^2_y were the largest compared to the other variance estimates. Moreover, their standard errors were very large. In the seed pool, the relative size of the estimate of σ^2_{gy} with respect to σ^2_g was higher in the CYC analysis than in the MY analysis. Thus, the estimate $\hat{\rho}$ obtained from the CYC analysis (0.213) was considerably smaller than from the MY analysis (0.582). In the pollen pool, the relative size of the estimate of σ^2_{gy} with respect to σ^2_g was smaller in the CYC analysis than in the MY analysis. Thus, the estimate $\hat{\rho}$ from the CYC analysis (0.578) was higher than from the MY analysis (0.498). Furthermore, the standard errors of $\hat{\rho}$ in the CYC analysis were slightly smaller than those in the MY analysis for both pools. The standard error of the estimate of ρ was computed via the delta method (Lynch and Walsh 1998; Ver Hoef 2012) as implemented in the `vpredict()` function of ASReml-R (Butler et al. 2017).

The asymptotic correlations between the GBV and GBV × year variance estimates by the CYC and MY analyses for both pools given in Table 4 are negative. The asymptotic correlation in the CYC analysis was higher than in the MY analysis in the seed pool, while in the

Table 3 Variance estimates and standard errors for year $\left(\sigma^2_y\right)$, GBV $\left(\sigma^2_g\right)$, GBV × year interaction effects $\left(\sigma^2_{gy}\right)$ in CYC and MY based on Eq. 2, and the $\hat{\rho}$ of the CYC and the MY analyses

Pool	Variance	MY		CYC	
		Estimate	SE	Estimate	SE
Seed pool					
	σ^2_y	19.796	16.071	25.275	17.661
	σ^2_g	1.960	0.203	1.279	0.242
	σ^2_{gy}	1.406	0.151	4.713	0.279
	$\hat{\rho}$	0.582	0.044	0.213	0.038
Pollen pool					
	σ^2_y	19.882	16.243	18.020	12.757
	σ^2_g	3.338	0.280	4.268	0.289
	σ^2_{gy}	3.367	0.221	3.120	0.187
	$\hat{\rho}$	0.498	0.032	0.578	0.026

SE standard error

Table 4 Asymptotic correlations between the GBV and GBV × year variance estimates by the CYC and MY analyses for both pools

Pool	Method	
	MY	CYC
Seed pool	−0.500	−0.565
Pollen pool	−0.469	−0.399

Table 5 Variance estimates of the apparent GBV $\left(\sigma^2_{\tilde{g}}\right)$, and the GBV $\left(\sigma^2_g\right)$, and the GBV × year $\left(\sigma^2_{gy}\right)$ interaction effects in GCA1-2020

Pool	Variance	First fit of GCA1-2020	After adjustment using $\hat{\rho}$ value	
			MY	CYC
Seed pool				
	$\sigma^2_{\tilde{g}}$	3.069	–	–
	σ^2_g	–	1.787	0.655
	σ^2_{gy}	–	1.282	2.414
Pollen pool				
	$\sigma^2_{\tilde{g}}$	4.418	–	–
	σ^2_g	–	2.199	2.552
	σ^2_{gy}	–	2.218	1.866

Table 6 The computing time for Eq. 2 in the GCA1 assessment on a desktop computer with an Intel i7 CPU, 64 GB RAM, and the Windows 10 (Version 21H2) operating system

Methods and datasets	Computing time	
	Seed pool	Pollen pool
MY	~2 h	~12 h
CYC	~2.3 h	~15 h
GCA1-2020 (first run)	~53 s	~7 min
GCA1-2020 (Rerun and obtain **C**)	~1.5 min	~10 min

Table 7 Mean correlation coefficient between the true genetic values ($\mathbf{g}$) and the GBLUPs $(\hat{\mathbf{g}})$, and mean correlation coefficient between $rank(\mathbf{g})$ and $rank(\hat{\mathbf{g}})$ in the GCA1 assessment for both pools

Pool	Correlations	Correlation estimate	
		MY	CYC
Seed pool	$r(\mathbf{g}, \hat{\mathbf{g}})$	0.688	0.420
	$r[rank(\mathbf{g}), rank(\hat{\mathbf{g}})]$	0.665	0.396
Pollen pool	$r(\mathbf{g}, \hat{\mathbf{g}})$	0.653	0.703
	$r[rank(\mathbf{g}), rank(\hat{\mathbf{g}})]$	0.631	0.682

pollen pool, the asymptotic correlation in the CYC analysis was smaller than the MY analysis. The pollen pool had smaller correlations than the seed pool, with the strongest correlation equal to −0.565. Thus, in general, there was mild confounding of effects between the GBV and GBV × year effects. The asymptotic correlations were computed from the inverse of the average information (AI) matrix of the variance and covariance estimates.

The variance component estimates of σ^2_{g}, σ^2_{g}, and σ^2_{gy} for the current year of both pools with each adjustment method are given in Table 5. The estimates of σ^2_{g} were obtained from the first fit of GCA1-2020 using Model 3, which were 3.069 and 4.418 for the seed and pollen pools, respectively. Then, using the $\hat{\rho}$ value from each MY and CYC methods, the estimates of σ^2_{g} and σ^2_{gy} were computed, and Model 3 was refitted using these estimates σ^2_{g} and σ^2_{gy}. The estimate of σ^2_{g} resulted in smaller values in the MY analysis for the pollen pool than with the CYC analysis, while in the seed pool, the estimates were on the opposite. Thus, the smaller $\hat{\rho}$ value in the CYC analysis in the seed pool led to a higher estimate of σ^2_{gy}.

Table 6 presents the computing time for all analyses for GCA1. As expected, with higher entry numbers, the computation time increased, as shown for the pollen pool. The computing time was more than 10 h for the pollen pool on a desktop computer with an Intel i7 CPU, 64 GB RAM, and the Windows 10 (Version 21H2) 64-bit operating system. This extensive computing time was due to the presence of GBV × year interaction effect directly in the model. The computing time for the current-year analysis (GCA1-2020) was considerably longer in the pollen pool than in the seed pool due to the higher number of entries. Moreover, the computing times for the simulations were marginal compared to the computing times for fitting the GP models.

The mean correlation coefficient between the true genetic values ($\mathbf{g}$) and the GBLUPs $(\hat{\mathbf{g}})$ and the mean correlation coefficient between $rank(\mathbf{g})$ and $rank(\hat{\mathbf{g}})$ of the GCA1 assessment for both pools based on 100 K simulations are presented in Table 7. Also, it is concordant with the breeder's objective to correctly rank entries. In the pollen pool, all correlation coefficients of the MY analysis were lower than the CYC analysis, while in the seed pool, all correlation coefficients of the CYC analysis were lower than the MY analysis. This is explained by the relative GBV × year $\left(\sigma^2_{\mathrm{gy}}\right)$ variance estimates compared to the GBV $\left(\sigma^2_{\mathrm{g}}\right)$ variance estimates (Table 7).

The quantity of interest from the simulation is depicted as a plot of the probabilities of obtaining truly best entries for each selected proportion of the entries based on the GBLUPs, as presented in Fig. 4. For a simulation with 100 K iterations, the seed pool took around 7 min and the pollen pool took around 17 min on a desktop computer with an Intel i7 CPU, 64 GB RAM, and the Windows 10 (Version 21H2) operating system. The probability plots for the CYC and MY analyses are noticeably different in the seed pool, in which the probability plot for MY shows curves approaching a probability of one faster compared to the CYC analysis. The discrepancy of the probability between MY and CYC in the seed pool was due to the much smaller value of $\hat{\rho}$ from the CYC analysis. Also, the correlation between true genetic values and GBLUPs was only 0.421 and so affected the shape of the curve for seed pool in the CYC analysis. In the CYC analysis, the probability of obtaining the 15 truly best entries when the number of selected entries (n) is 50% of N was only around 0.21, while with the MY analysis, the probability was around 0.80. In the pollen pool, both MY and CYC showed a similar trend. Both analyses implied that by selecting a small proportion of entries, i.e. 30%, the probability of obtaining truly ten best entries was around 0.76 and 0.62 for CYC and MY analyses, respectively. Moreover, Fig. 4 agrees with the correlation coefficients in Table 8, in that the higher correlation between the true genetic values and the GBLUPs, the higher the probability achieved by selecting a smaller number of entries.

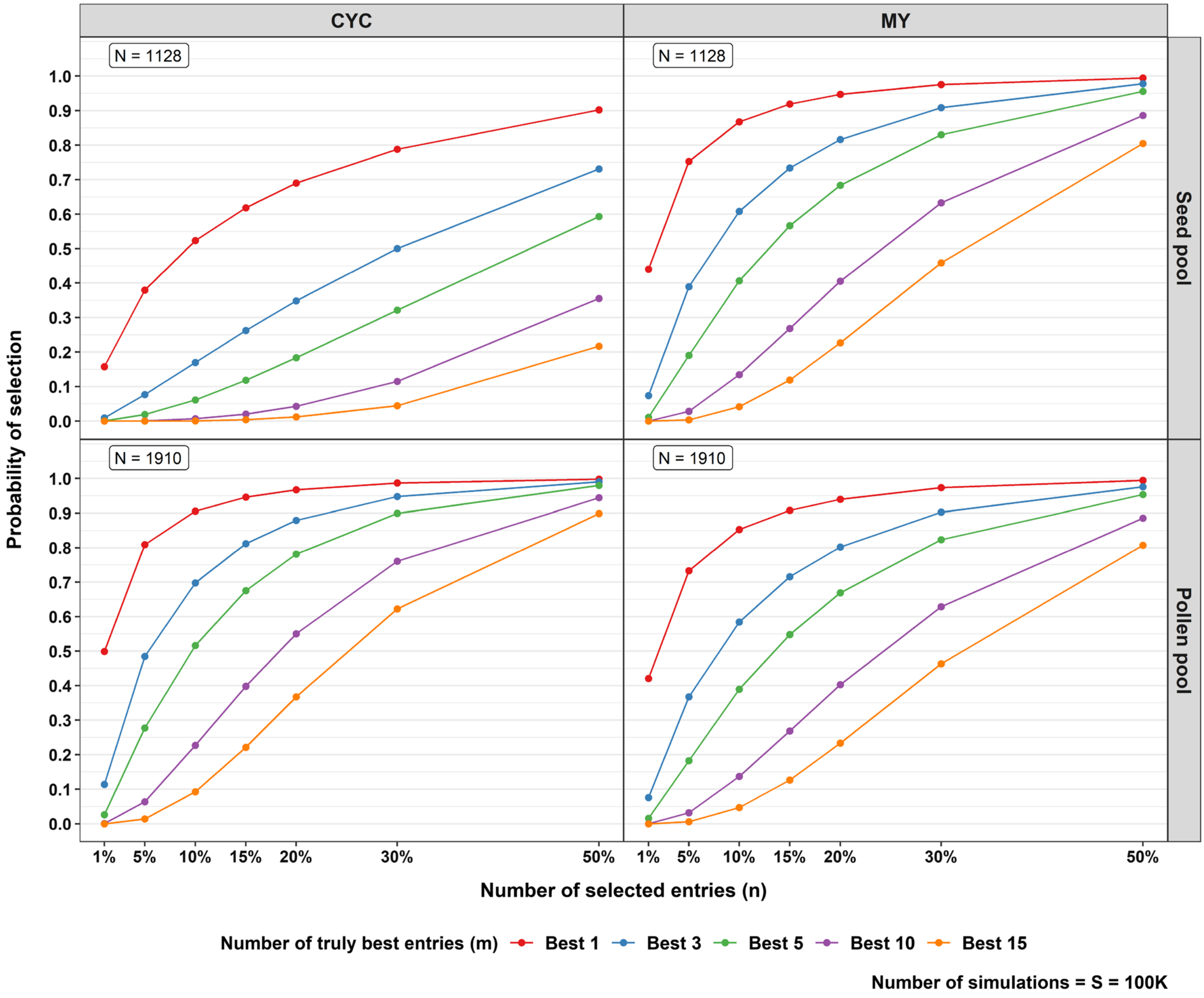

Fig. 4 Plots of the probability of obtaining the *m* truly best entries based on the GBLUPs for each selected number of entries (expressed as percentage of *N*) from GCA1 assessment of each pool. The different coloured entries indicate the different numbers (*m*) of truly best entries

Table 8 Variance estimates and standard errors for year $\left(\sigma_y^2\right)$, GBV $\left(\sigma_g^2\right)$, GBV × year interaction effects $\left(\sigma_{gy}^2\right)$ for each cycle of the GCA2 assessment

Pool	Variance	Estimate and standard error (SE)							
		Cycle 1		Cycle 2		Cycle 3		Cycle 4	
		Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Seed pool									
	σ_y^2	9.236	25.546	6.699	45.889	0.000	–	0.000	–
	σ_g^2	0.634	0.321	0.697	0.228	1.306	0.338	0.320	0.212
	σ_{gy}^2	0.486	0.251	0.086	0.125	0.000	–	0.247	0.187
Pollen pool									
	σ_y^2	0.000	–	15.696	56.278	0.000	–	0.000	–
	σ_g^2	4.269	0.754	0.707	0.206	1.707	0.388	0.396	0.165
	σ_{gy}^2	0.780	0.305	0.136	0.111	0.452	0.200	0.193	0.141

SE standard error

GCA2 response to selection assessment

The variance estimates and standard errors for year $\left(\sigma^2_{\mathrm{y}}\right)$, GBV $\left(\sigma^2_{\mathrm{g}}\right)$, and GBV × year interaction $\left(\sigma^2_{\mathrm{gy}}\right)$ effects for each cycle are given in Table 8. In Cycle 4, the σ^2_{g} estimates were the smallest and had the smaller ratio with the σ^2_{gy} estimates for both pools. In Cycles 1 to 3, both pools had a relatively higher ratio of σ^2_{g} and σ^2_{gy} estimates.

In the seed pool, the year variance estimates in Cycle 3 to Cycle 4 were zero, while in the pollen pool, it was in the Cycle 1, Cycle 3, and Cycle 4. Furthermore, in the seed pool, the σ^2_{gy} estimate in Cycle 3 was also zero. On the other hand, the year variance estimate in Cycle 2 of the pollen pool was relatively large, i.e. 15.696. Furthermore, the standard errors for the year variance estimates were large for both pools. In general, when a variance estimate goes to zero, this is possibly biased due to a small sample size. In our study, the number of years was only two years. Thus, this might be the reason that the year variance estimates were mostly zero.

The asymptotic correlations between the GBV and GBV × year variance estimates of each cycle for both pools presented in Table 9 are negative with values ranging from −0.21 to −0.49. However, there was a trend that the asymptotic correlations increased from Cycle 1 to Cycle 4 in both pools. The asymptotic correlations of the seed pool were mostly higher than the pollen pool. The highest asymptotic correlation was −0.493 in the Cycle 4 of the seed pool. In the same cycle, the asymptotic correlation for the pollen pool was −0.382. In the Cycle 3 of the seed pool, the asymptotic correlation was zero due to the variance estimate for GBV × year was zero. So, generally, there was only mild confounding between the GBV and GBV × year effects.

Table 10 presents the computing time of each cycle analysis for each pool. Due to relatively much smaller number of entries, the seed pool took around 2 to 5 s, while in the pollen pool, it took around 13 to 15 s. Furthermore, the GCA2 analyses were done using the single-stage approach, which was feasible due to a small number of entries and the number of years being only two.

Table 11 provides the mean correlation coefficients between true genetic values ($\mathbf{g}$) and the GBLUPs $\left(\hat{\mathbf{g}}\right)$, and the mean correlation coefficients between $rank(\mathbf{g})$ and $rank\left(\hat{\mathbf{g}}\right)$ of each cycle of the GCA2 assessment for both pools based on 100 K simulations. The lowest correlation coefficients for both pools were observed in Cycle 4 due to smaller σ^2_{g} estimates and their ratio to the σ^2_{gy} estimates compared to the other cycles. The correlation coefficient between $\mathbf{g}$ and $\hat{\mathbf{g}}$ was 0.616 and 0.673 for the seed and pollen pools, respectively. These coefficients were much lower than the correlation coefficients in other Cycles in both pools. The pattern of the correlation coefficients of the rank of $\mathbf{g}$ and $\hat{\mathbf{g}}$ was the same as the correlation coefficients between $\mathbf{g}$ and $\hat{\mathbf{g}}$.

Table 9 Asymptotic correlations between the GBV and GBV × year variance estimates of each cycle of the GCA2 assessment for both pools

Pool	Asymptotic correlation			
	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Seed pool	−0.400	−0.268	0.000	−0.493
Pollen pool	−0.205	−0.307	−0.333	−0.382

Table 10 The computing time for all analyses in the GCA2 assessment on a desktop computer with an Intel i7 CPU, 64 GB RAM, and the Windows 10 (Version 21H2) operating system

Pool	Computing time			
	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Seed pool	~2 s	~4 s	~5 s	~4 s
Pollen pool	~13 s	~13 s	~15 s	~15 s

The quantity of interest from the simulation for GCA2 assessment is depicted as a probability plot of obtaining truly best entries given the number of selected entries in Fig. 5. The 100 K simulations took around 2 min for each cycle in each pool on a desktop computer with an Intel i7 CPU, 64 GB RAM, and the Windows 10 (Version 21H2) operating system. As we can see, the differing number of entries played a significant role. The pollen pool had a relatively larger number of entries compared to the seed pool. The probability of obtaining the truly best entry was, therefore, higher, especially in Cycle 1. In Cycle 1, due to a much lower number of entries in the seed pool, selecting 20% entry results in a probability of nearly 0 of having picked the ten truly best entries, while in the pollen pool, the probability was around 0.6. In Cycle 4, although the number of entries in both pools was decent, both pools had relatively low ratios of the GBV and GBV × year variance estimates as shown in Table 9, and so the correlations

Table 11 Mean correlation coefficient between the true genetic values ($\mathbf{g}$) and the GBLUPs $\left(\hat{\mathbf{g}}\right)$, and mean correlation coefficient between $rank(\mathbf{g})$ and $rank\left(\hat{\mathbf{g}}\right)$ in the GCA2 assessment for both pools

Pool	Correlations	Correlation estimate			
		Cycle 1	Cycle 2	Cycle 3	Cycle 4
Seed pool	$r\left(\mathbf{g}, \hat{\mathbf{g}}\right)$	0.736	0.820	0.865	0.616
	$r\left[rank(\mathbf{g}), rank\left(\hat{\mathbf{g}}\right)\right]$	0.700	0.790	0.839	0.578
Pollen pool	$r\left(\mathbf{g}, \hat{\mathbf{g}}\right)$	0.899	0.804	0.837	0.673
	$r\left[rank(\mathbf{g}), rank\left(\hat{\mathbf{g}}\right)\right]$	0.883	0.778	0.816	0.645

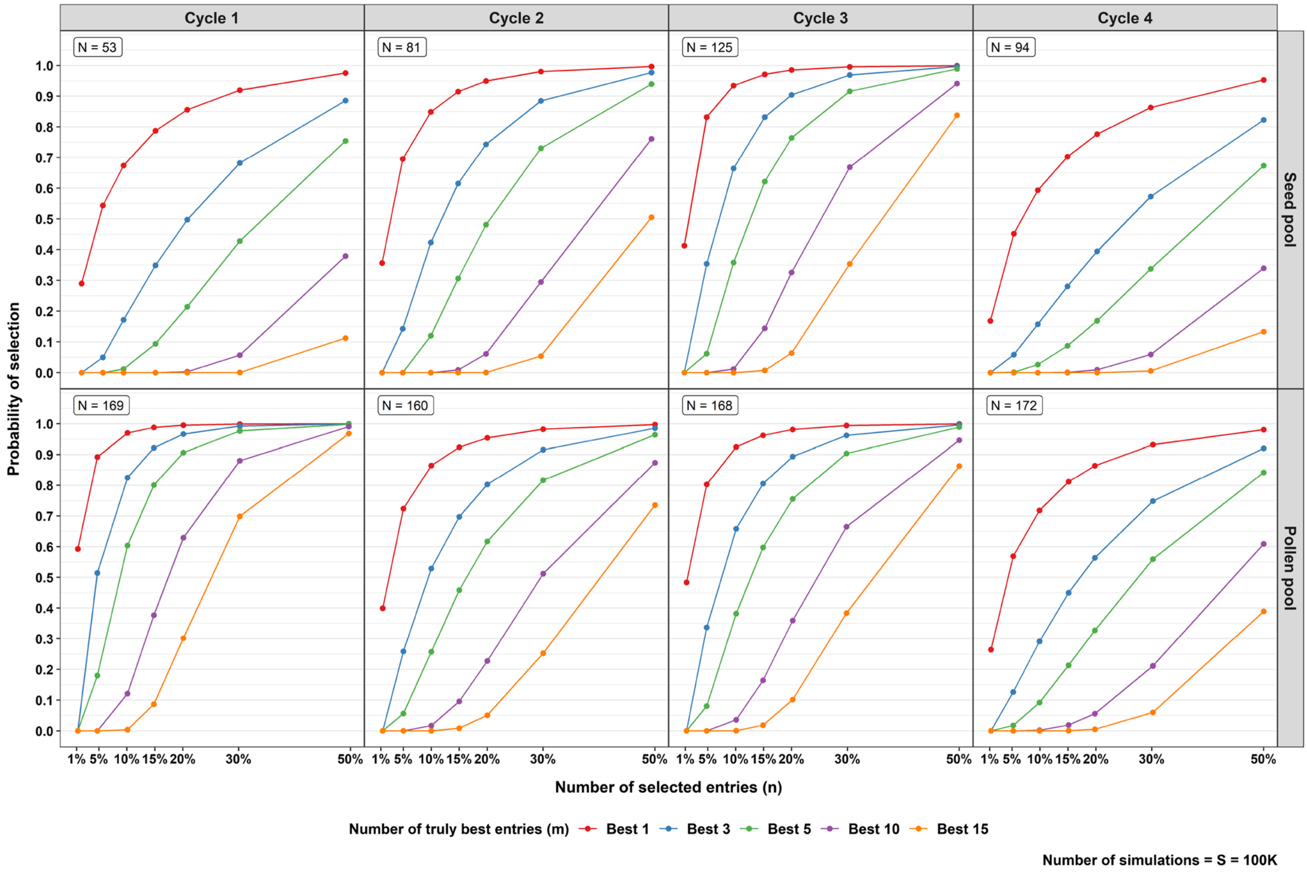

Fig. 5 Plots of the probability of obtaining the *m* truly best entries based on the GBLUPs for each selected number of entries (expressed as percentage of *N*) of each cycle in the GCA2 assessment for each pool. The different coloured entries indicate the different numbers (*m*) of truly best entries. In general, the pollen pool has higher probability to obtain the truly best entries than the seed pool. In Cycle 4, both pools have a relatively lower probability to obtain the truly best entries compared to the other cycles

between true GBV and GBLUPs were relatively low, as shown in Table 12. Thus, the probability of obtaining truly best entries was relatively low compared to Cycles 2 and 3. For example, having the ten truly best entries by selecting 20% of the entries only achieved a probability of around 0.06 for the pollen pool and nearly 0 for the seed pool.

Discussion

The response to selection has been widely used in plant breeding programs to measure genetic gain. Our proposed simulation approach allows breeders to obtain information on how well genomic selection can be made for the next stage. The response to selection can be measured in terms of the probability of selecting the truly best entry based on the number or proportion of selected entries. Knowing the probability of selecting the best *m* lines based on the proportion of the selected lines allows breeders to take decisions as to modifications of selection fractions in selection stages, trials, and years with increased error variances of mean estimates due to unfavourable environmental conditions (e.g. drought) to ensure that a certain portion of truly best genotypes are continued to the next selection stages. Likewise, in labour-intensive trait introgression (conversion) procedures with large numbers of entries in early selection stages, this probability allows optimizing the allocation of technical and personnel resources through an estimate of the risk of potential loss of superior lines in dependence of the number of the total number of converted entries.

In the GCA1, the number of entries is normally large since this is the first year that all the entries enter the GCA testing trial after a *per-se* line selection stage for agronomic (non-yield) traits. In comparison with using only GCA1 from the years 2016 to 2019 as in the MY analysis, the CYC analysis benefits from using the entries in the selected fractions of GCA2 and GCA3 improving the connectivity across GCA trials. In the same vein, Smith et al. (2021) also reported that the poor connectivity could lead to poor estimated genetic variance parameters, and so decreasing the genetic gain (Sales and Hill 1976a, b). Furthermore, the use of the kinship matrix to model GBV × year interactions can be advantageous, as Bernal-Vasquez et al. (2017) reported. By using the kinship, we gained more connectivity across years. This is because, even when the entries were only available in one year, the same marker alleles are assessed across years.

Discrepancies between MY and CYC in the pollen pool were less pronounced compared to the seed pool. The connectivity in the CYC analysis slightly improved the estimate of ρ in the pollen pool but not in the seed pool, which might be due to the genetic sampling in the seed pool population. In the seed pool, the CYC analysis had higher probability of bias because, although the GCA2 and GCA3 entries improve connectivity, the selected fraction may be composed of lines that trace back in their pedigree to only very few parental components (pre-dominant family selection) that show a very particular change of ranking between years that leads to an over-estimation of GBV × year and an underestimation of ρ. Furthermore, Table 4 shows that the GBV variance estimates of the seed pool are far smaller compared to the pollen pool in both MY and CYC. The risk of such a bias is expected to be potentiated when the environmental conditions change considerably between years. Here, the year of 2017 was extremely wet, whereas the years of 2018, 2019, and 2020 were very dry. Such effects can lead to biases in the estimates of relative GBV × year estimates based on the effects of pre-dominant families.

We have also shown that the simulation for the GCA2 assessment is useful to measure the selection accuracy for the final trial in the GCA3. Thus, the simulation was only based on the selected entries that progressed to GCA2.

In this study, the number of years was not large, i.e. only 4 years in the GCA1 assessment, and the simulations were conducted using the estimates produced by frequentist residual maximum likelihood method (REML). Furthermore, as demonstrated by our results, estimation of the variance components for GBV and GBV × year effects is the Achilles' Heel of the whole approach. These variance components are expected to display year-to-year variation, which is why our approach prescribes re-estimation of each variance in a new year using a long-term estimate of ρ. The expected year-to-year variability, however, suggests that a fully Bayesian approach that operates on a prior distribution for each variance component, rather than a fixed value, may be beneficial. The key challenge with such a framework is how to properly inform these priors and the need to integrate long-term data from an ongoing breeding program.

An approach that can be beneficial is to continuously use a Bayesian framework to collect more information from the previous years and use it as a prior information to update the current-year analysis. This approach is known as Bayesian updating (Sorensen and Gianola 2002). In Bayesian updating, the prior distribution is based on the previous posterior distribution. In this case, the Bayes theorem has "memory", and the inferences can be updated sequentially. For example, for the GCA1 analysis, the prior distribution for the current-year analysis can be obtained from the posterior distribution of many previous GCA1 or Cycle analyses. Therefore, a further study with the Bayesian updating framework would be worthwhile to investigate in the future.

Furthermore, the Gibbs sampling for estimating variance components might be appealing compared to REML since it used prior distribution that could produce more accurate variance estimates (Van Tassell et al. 1995). In the GCA2 assessment, we only chose the full set of entries that progressed to GCA2 from GCA1, and so we did not face the missing-not-at-random pattern that can lead to a bias by using REML (Piepho and Möhring, 2006; Hartung and Piepho, 2021). However, the year variance estimates were mostly zero in the GCA2 assessment, resulting from a non-negativity constraint on REML estimates. In this study, the zero estimates of the year variance might be due to a small number of years. Modelling the year effect as fixed can be an alternative, as shown in Table S.1 in Supplementary Tables. In Table S.1, the GBV and GBV × year variance estimates are very close to their estimates in Table 4. Thus, the estimates of ρ in Table S.2 were also very similar to the estimates in Table 4. A similar pattern also was shown for the GCA2 assessment. The variance component estimates of GBV and GBV × year in Table 9 and in Table S.3 of Supplementary Tables are very similar. Thus, modelling the year effect as fixed can be an alternative option.

Despite issues around the best dataset for estimating the variance components, we demonstrated that the genetic gain and the probabilities of selection of superior candidate lines can be measured through simulations in GP framework, either using CYC or MY analyses. We recommend choosing the analysis based on the breeding population structure. The MY analysis is suitable when the population structure is highly influenced by strong selection intensity, e.g. the seed pool, while the CYC analysis can be chosen when there is no pre-dominant selection, e.g. the pollen pool. A further comparative study to compare the MY and CYC approaches using other crops and breeding populations is, therefore, worthwhile to be conducted.

Appendix

See Tables 12, 13, and 14.

Table 12 The variance–covariance structures for each term in Eqs. 1 and 3

Term and assumption	Variance–covariance	Remarks
$\mathbf{u}_{\mathbf{l}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{l}})$	$\mathbf{G}_{\mathbf{l}} = \mathbf{I}\sigma^2_{\mathrm{l}}$	Identity variance structure for each random effect
$\mathbf{u}_{\mathbf{t}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{t}})$	$\mathbf{G}_{\mathbf{t}} = \mathbf{I}\sigma^2_{\mathrm{t}}$	
$\mathbf{u}_{\mathbf{gt}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{gt}})$	$\mathbf{G}_{\mathbf{gt}} = \mathbf{I}\sigma^2_{\mathrm{gt}}$	
$\mathbf{u}_{\mathbf{gl}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{gl}})$	$\mathbf{G}_{\mathbf{gl}} = \mathbf{I}\sigma^2_{\mathrm{gl}}$	
$\mathbf{u}_{\mathbf{ls}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{ls}})$	$\mathbf{G}_{\mathbf{ls}} = \mathbf{I}\sigma^2_{\mathrm{ls}}$	
$\mathbf{u}_{\mathbf{lsr}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{lsr}})$	$\mathbf{G}_{\mathbf{lsr}} = \mathbf{I}\sigma^2_{\mathrm{lsr}}$	
$\mathbf{u}_{\mathbf{lsrb}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{lsrb}})$	$\mathbf{G}_{\mathbf{lsrb}} = \mathbf{I}\sigma^2_{\mathrm{lsrb}}$	
$\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$	$\mathbf{R} = \oplus_{l=1}^{L}\mathbf{R}_l$ $\mathbf{R}_l = \mathbf{I}\sigma^2_{\varepsilon_l}$	Heterogeneous location-specific consisting of a block-diagonal matrix with diagonal element $\sigma^2_{\varepsilon_l}, l = 1, 2, \ldots, L$

The subscript l denotes the lth location; L is the number of locations

Table 13 The variance–covariance structures for each term in Eqs. 2 and 3

Term and assumption	Variance–covariance	Remarks
$\mathbf{u}_{\mathbf{g}} \sim MVN(\mathbf{0}, \mathbf{K}\sigma^2_{\mathrm{g}})$	$\mathbf{K}\sigma^2_{\mathrm{g}}$	$\mathbf{u}_{\mathbf{g}} = \mathbf{Qv}$, where $\mathbf{Q}$, is the marker genotypes matrix, $\mathbf{v}$ is the vector of marker effects, $\mathbf{v} \sim N(0, \mathbf{I}\sigma^2_{\mathrm{g}})$, where σ^2_{g} is the genomic variance. Based on these assumptions $\mathbf{u}_{\mathbf{g}} \sim MVN(0, \mathbf{K}\sigma^2_{\mathrm{g}})$, where $\mathbf{K} = \mathbf{QQ}^{\mathbf{T}}$ The $\mathbf{K}$ matrix is the genomic relationship
$\mathbf{u}_{\mathbf{gy}} \sim MVN(\mathbf{0}, \mathbf{G}_{\mathbf{gy}})^{\dagger}$	$\mathbf{G}_{\mathrm{gy}} = \oplus_{j=1}^{J}\mathbf{K}_j\sigma^2_{\mathrm{gy}}$	The matrix $\mathbf{Z}_{\mathbf{gy}}$ is a block-diagonal matrix with blocks given by the coefficient of entries in a given year, $\oplus_{j=1}^{J}\mathbf{Z}_{\mathbf{gy}_j}$, where j is the jth year $\left(\mathbf{Z}_{\mathbf{gy}_j}\right)$, $\mathbf{Z}_{\mathbf{gy}} = \begin{bmatrix} \mathbf{Z}_{\mathbf{gy}_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{\mathbf{gy}_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{\mathbf{gy}_j} \end{bmatrix}$ and $\mathrm{var}\left(\mathbf{u}_{\mathbf{gy}}\right) = \mathbf{G}_{\mathbf{gy}}$, where $\mathbf{G}_{\mathbf{gy}} = \oplus_{j=1}^{J}\mathbf{K}_j\sigma^2_{\mathrm{gy}}$, $\mathbf{K}_j$ is the kinship of all entries tested in the jth year

The subscript j denotes the jth year; J is the number of years.

[†]Only applies to Eq. 2

Table 14 The variance–covariance structures for each term in Eq. 4 for the GCA2 assessment

Term and assumption	Variance–covariance	Remarks
$\mathbf{u}_{\mathbf{y}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{y}})$	$\mathbf{G}_{\mathbf{y}} = \sigma^2_{\mathrm{y}}$	Identity variance structure
$\mathbf{u}_{\mathbf{l}(\mathrm{y})} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{l}})$	$\mathbf{G}_{\mathbf{l}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{l}(j)}$ $\mathbf{G}_{\mathbf{l}(j)} = \mathbf{I}\sigma^2_{\mathrm{l}_{(j)}}$	Heterogeneous year-specific variance structure in which the diagonal elements $\sigma^2_{\mathrm{l}_{(j)}}$, $\sigma^2_{\mathrm{t}_{(j)}}$, $\sigma^2_{\mathrm{gt}_{(j)}}$, $\sigma^2_{\mathrm{gl}_{(j)}}$, $\sigma^2_{\mathrm{ls}_{(j)}}$, $\sigma^2_{\mathrm{lsr}_{(j)}}$, $\sigma^2_{\mathrm{lsrb}_{(j)}}$ differ for jth year for each $\mathbf{G}_{\mathbf{l}(j)}$, $\mathbf{G}_{\mathbf{t}(j)}$,$\mathbf{G}_{\mathbf{gt}(j)}$, $\mathbf{G}_{\mathbf{gl}(j)}$, $\mathbf{G}_{\mathbf{ls}(j)}$, $\mathbf{G}_{\mathbf{lsr}(j)}$, $\mathbf{G}_{\mathbf{lsrb}(j)}$, respectively
$\mathbf{u}_{\mathbf{t}_{(\mathrm{y})}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{t}})$	$\mathbf{G}_{\mathbf{t}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{t}(j)}$ $\mathbf{G}_{\mathbf{t}(j)} = \mathbf{I}\sigma^2_{\mathrm{t}_{(j)}}$	
$\mathbf{u}_{\mathbf{gt}_{(\mathrm{y})}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{gt}})$	$\mathbf{G}_{\mathbf{gt}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{gt}(j)}$ $\mathbf{G}_{\mathbf{gt}(j)} = \mathbf{I}\sigma^2_{\mathrm{gt}_{(j)}}$	
$\mathbf{u}_{\mathbf{gl}_{(\mathrm{y})}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{gl}})$	$\mathbf{G}_{\mathbf{gl}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{gl}(j)}$ $\mathbf{G}_{\mathbf{gl}(j)} = \mathbf{I}\sigma^2_{\mathrm{gl}_{(j)}}$	
$\mathbf{u}_{\mathbf{ls}_{(\mathrm{y})}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{ls}})$	$\mathbf{G}_{\mathbf{ls}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{ls}(j)}$ $\mathbf{G}_{\mathbf{ls}(j)} = \mathbf{I}\sigma^2_{\mathrm{ls}_{(j)}}$	
$\mathbf{u}_{\mathbf{lsr}_{(\mathrm{y})}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{lsr}})$	$\mathbf{G}_{\mathbf{lsr}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{lsr}(j)}$ $\mathbf{G}_{\mathbf{lsr}(j)} = \mathbf{I}\sigma^2_{\mathrm{lsr}_{(j)}}$	
$\mathbf{u}_{\mathbf{lsrb}_{(\mathrm{y})}} \sim N(\mathbf{0}, \mathbf{G}_{\mathbf{lsrb}})$	$\mathbf{G}_{\mathbf{lsrb}} = \oplus_{j=1}^{J} \mathbf{G}_{\mathbf{lsrb}(j)}$ $\mathbf{G}_{\mathbf{l}}\boldsymbol{srb}_{(j)} = \mathbf{I}\sigma^2_{\mathrm{lsrb}_{(j)}}$	
$\mathbf{u}_{\mathbf{g}} \sim MVN(\mathbf{0}, \mathbf{K}\sigma^2_{\mathrm{g}})$	$\mathbf{K}\sigma^2_{\mathrm{g}}$	$\mathbf{u}_{\mathbf{g}} = \mathbf{Qv}$, where $\mathbf{Q}$ is the marker genotypes matrix, $\mathbf{v}$ is the vector of marker effects, $\mathbf{v} \sim N(0, \mathbf{I}\sigma^2_{\mathrm{g}})$, where σ^2_{g} is the genomic variance. Based on these assumptions $\mathbf{u}_{\mathbf{g}} \sim MVN(0, \mathbf{K}\sigma^2_{\mathrm{g}})$, where $\mathbf{K} = \mathbf{QQ}^{\mathbf{T}}$ The $\mathbf{K}$ matrix is the genomic relationship
$\mathbf{u}_{\mathbf{gy}} \sim MVN(\mathbf{0}, \mathbf{G}_{\mathbf{gy}})$	$\mathbf{G}_{\mathbf{gy}} = \oplus_{j=1}^{2} \mathbf{K}_j \sigma^2_{\mathrm{gy}}$	The matrix $\mathbf{Z}_{\mathbf{gy}}$ is a block-diagonal matrix with blocks given by the coefficient of entries in a given year, $\oplus_{j=1}^{2} \mathbf{Z}_{\mathbf{gy}_j}$, $\mathbf{Z}_{\mathbf{gy}} = \begin{bmatrix} \mathbf{Z}_{\mathbf{gy}_1} & 0 \\ 0 & \mathbf{Z}_{\mathbf{gy}_2} \end{bmatrix}$ and $\operatorname{var}\left(\mathbf{u}_{\mathbf{gy}}\right) = \mathbf{G}_{\mathbf{gy}}$, where $\mathbf{G}_{\mathbf{gy}} = \oplus_{j=1}^{2} \mathbf{K}_j \sigma^2_{\mathrm{gy}}$, $\mathbf{K}_j$ is the kinship of all entries tested in the jth year
$\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$	$\mathbf{R} = \oplus_{j=1}^{J} \mathbf{R}_{l_{(j)}}$ $\mathbf{R}_{l_{(j)}} = \mathbf{I}\sigma^2_{\varepsilon_{(j)}}$	Heterogeneous year × location-specific variance structure consisting of a block-diagonal matrix with diagonal elements $\sigma^2_{\varepsilon_{(j)}}$, $j = 1, 2, \ldots, J; l = 1, 2, \ldots, L$

The subscript j denotes the jth year and subscript l denotes the lth location. The J is the number of years and L is the number of locations

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-022-04157-1.

Acknowledgements The authors thank the editor in charge and anonymous reviewers for insightful comments that helped us improve the manuscript substantially.

Author contribution statement HPP, AG, VM, and AMBV planned and designed the research project. HB wrote the R codes, performed the analysis, and wrote the manuscript. AG, AMBV, VM, and MS provided the scientific advice. All the authors read, edited, and approved the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Code availability The R codes are available as electronic Supplementary Materials.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval The authors declare that the experiments comply with the current laws of Germany.

References

Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471. https://doi.org/10.1214/09-STS307

Bauer E et al (2017) Towards a whole-genome sequence for rye (*Secale cereale* L.). Plant J 89:853–869. https://doi.org/10.1111/tpj.13436

Bernal-Vasquez A-M, Gordillo A, Schmidt M, Piepho H-P (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC Genet 18:51. https://doi.org/10.1186/s12863-017-0512-8

Buntaran H, Piepho H-P, Schmidt P, Rydén J, Halling M, Forkman J (2020) Cross-validation of stagewise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Sci 60:2221–2240. https://doi.org/10.1002/csc2.20177

Butler DG, Cullis B, Gilmour A, Gogel BJ, Thompson R (2017) ASReml-R reference manual, version 4. University of Wollongong, Wollongong

Crossa 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

Endelman JB, Jannink J-L (2012) Shrinkage estimation of the realized relationship matrix. Genes Genomes Genetics (G3) 2:1405–1413. https://doi.org/10.1534/g3.112.004259

Feldmann MJ, Piepho H-P, Bridges WC, Knapp SJ (2020) Accurate estimation of marker-associated genetic variance and heritability in complex trait analyses. bioRxiv. https://doi.org/10.1101/2020.04.08.032672

Gezan SA, de Oliveira AA, Murray D (2021) ASRgenomics: an R package with complementary genomic functions. VSN International, Hemel Hempstead

Hartung J, Piepho H-P (2021) Effect of missing values in multi-environmental trials on variance component estimates. Crop Sci 61:4087–4097. https://doi.org/10.1002/csc2.20621

Jarquín D et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoret Appl Genet 127:595–607. https://doi.org/10.1007/s00122-013-2243-1

Kleinknecht K, Möhring J, Laidig F, Meyer U, Piepho HP (2016) A simulation-based approach for evaluating the efficiency of multi-environment trial designs. Crop Sci 56:2237–2250

Lorenz AJ et al (2011) Genomic selection in plant breeding: knowledge and prospects. In: Sparks DL (ed) Advances in agronomy, vol 110. Academic Press, San Diego, pp 77–123

Lush JL (1942) Animal breeding plans, 2nd edn. The Iowa state College Press, Ames

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland

Martis MM et al (2013) Reticulate evolution of the rye genome. Plant Cell 25:3685–3698. https://doi.org/10.1105/tpc.113.114553

McLean RA, Sanders WL, Stroup WW (1991) A unified approach to mixed linear models. Am Stat 45:54–64. https://doi.org/10.1080/00031305.1991.10475767

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Patterson HD, Williams ER (1976) A new class of resolvable incomplete block designs. Biometrika 63:83–92

Piepho H-P, Möhring J (2006) Selection in cultivar trials—is it ignorable? Crop Sci 46:192–201. https://doi.org/10.2135/cropsci2005.04-0038

Piepho H-P, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics 177:1881–1888. https://doi.org/10.1534/genetics.107.074229

Piepho H-P, van Eeuwijk FA (2002) Stability analysis in crop performance evaluation. In: Kang MS (ed) Crop improvement: Challenges in the twenty-first century. The Haworth Press, New York, pp 315–351

Piepho HP, Ogutu JO, Schulz-Streeck T, Estaghvirou B, Gordillo A, Technow F (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci 52:1093–1104. https://doi.org/10.2135/cropsci2011.11.0592

R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

R Studio Team (2021) RStudio: Integrated development environment for R. RStudio, PBC, Boston

Rutkoski JE (2019) A practical guide to genetic gain. In: Sparks DL (ed) Advances in agronomy, vol 157. Academic Press, San Diego, pp 217–249. https://doi.org/10.1016/bs.agron.2019.05.001

Sales J, Hill WG (1976a) Effect of sampling errors on efficiency of selection indices 1. Use of information from relatives for single trait improvement. Anim Sci 22:1–17. https://doi.org/10.1017/S0003356100035364

Sales J, Hill WG (1976b) Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. Anim Sci 23:1–14. https://doi.org/10.1017/S0003356100031020

Smith A, Cullis B, Gilmour A (2001) Applications: The analysis of crop variety evaluation data in Australia. Aust N Z J Stat 43:129–145. https://doi.org/10.1111/1467-842x.00163

Smith A, Ganesalingam A, Lisle C, Kadkol G, Hobson K, Cullis B (2021) Use of contemporary groups in the construction of multi-environment trial datasets for selection in plant breeding programs. Front Plant Sci. https://doi.org/10.3389/fpls.2020.623586

Sorensen D, Gianola D (2002) Bayesian updating. In: Sorensen D, Gianola D (eds) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York, pp 249–257. https://doi.org/10.1007/0-387-22764-4_5

van Tassell CP, Casella G, Pollak EJ (1995) Effects of selection on estimates of variance components using gibbs sampling and restricted maximum likelihood. J Dairy Sci 78:678–692

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

Ver Hoef JM (2012) Who invented the Delta method? Am Stat 66:124–127. https://doi.org/10.1080/00031305.2012.687494

Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer, New York

Williams ER, Matheson AC, Harwood CE (2002) Experimental design and analysis for tree improvement, 2nd edn. CSIRO Publishing, Collingwood

Yang J et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569. https://doi.org/10.1038/ng.608