Theoretical and Applied Genetics (2024) 137:104
https://doi.org/10.1007/s00122-024-04592-2

ORIGINAL ARTICLE

Optimizing selection based on BLUPs or BLUEs in multiple sets of genotypes differing in their population parameters

Albrecht E. Melchinger[1,2] · Rohan Fernando[3] · Andreas J. Melchinger[4] · Chris-Carolin Schön[1]

Received: 27 October 2023 / Accepted: 5 March 2024

Abstract

Key message **Selection response in truncation selection across multiple sets of candidates hinges on their post-selection proportions, which can deviate grossly from their initial proportions. For BLUPs, using a uniform threshold for all candidates maximizes the selection response, irrespective of differences in population parameters.**

Abstract Plant breeding programs typically involve multiple families from either the same or different populations, varying in means, genetic variances and prediction accuracy of BLUPs or BLUEs for true genetic values (TGVs) of candidates. We extend the classical breeder's equation for truncation selection from single to multiple sets of genotypes, indicating that the expected overall selection response (ΔG_{Tot}) for TGVs depends on the selection response within individual sets and their post-selection proportions. For BLUEs, we show that maximizing ΔG_{Tot} requires thresholds optimally tailored for each set, contingent on their population parameters. For BLUPs, we prove that ΔG_{Tot} is maximized by applying a uniform threshold across all candidates from all sets. We provide explicit formulas for the origin of the selected candidates from different sets and show that their proportions before and after selection can differ substantially, especially for sets with inferior properties and low proportion. We discuss implications of these results for (a) optimum allocation of resources to training and prediction sets and (b) the need to counteract narrowing the genetic variation under genomic selection. For genomic selection of hybrids based on BLUPs of GCA of their parent lines, selecting distinct proportions in the two parent populations can be advantageous, if these differ substantially in the variance and/or prediction accuracy of GCA. Our study sheds light on the complex interplay of selection thresholds and population parameters for the selection response in plant breeding programs, offering insights into the effective resource management and prudent application of genomic selection for improved crop development.

Introduction

Selection is one of the major drivers of evolution and breeding. In nature, various types of selection occur, which are studied in evolutionary biology and described in textbooks on population genetics (e.g., Hartl et al. 1997). In breeding, directional selection is by far the most important type of selection in the sense that breeders typically select only a certain number or proportion of top candidates for a single trait or an index of the most important traits. The selected candidates are then advanced for further breeding or utilized as experimental cultivars for commercial purposes.

Cochran (1951) derived the primary mathematical results for the changes in population parameters under truncation selection in a seminal paper and demonstrated its application to plant selection. He described the selection response for a target variable, when selection is based on correlated variates. Cochran's formula and its extension to the peculiarities

Communicated by Antonio Augusto Franco Garcia.

✉ Albrecht E. Melchinger
albrechtmelchinger@gmail.com

[1] Plant Breeding, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany

[2] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany

[3] Department of Animal Science, Iowa State University, Ames, IA 50011, USA

[4] Department of Mathematics, University of Stuttgart, 70569 Stuttgart, Germany

Published online: 15 April 2024

in plant breeding, such as length of the breeding cycle and parental control, are known as the breeders' equation (cf. Bernardo 2002; Lynch and Walsh 1998). This equation is one of the most important contributions of quantitative genetics to practical breeding as it quantifies the relevant factors that determine the progress expected from directional selection. However, the breeders' equation strictly applies only to selection in a single population and assumes homogeneous correlation between the true genetic value (TGV) and the selection criterion (SC) for all candidates, which is generally not met in practice. More general settings, dropping the latter assumption, were investigated by Bulmer (1980).

In animal breeding, the problem of heterogeneity of variances among sets was early addressed in the context of different environmental groups (Brotherstone and Hill 1986). Hill (1984) found that under more intense selection, more animals are selected from the group with larger variance and recommended to correct for heterogeneity. For selection based on BLUPs, Garrick and Van Vleck (1987) examined the case of heterogeneous variances and showed that selection assuming homogeneity is still highly efficient if the prediction accuracy is high.

Plant breeding programs typically involve multiple sets of candidates from various families or populations (e.g., Auinger et al. 2021; Lian et al. 2014) and breeders often apply the same threshold to all candidates without considering their origin. However, if the sets differ in their mean and/or genetic variance and/or heritability (h^2) of entry means, calculated as best linear unbiased estimates (BLUEs) in phenotypic selection, this may be suboptimal for the selection response of the entire program. This problem arises for example when one set of candidates is tested in more locations and/or years than another set, resulting in different heritabilities (h^2).

When selection is based on best linear unbiased predictors (BLUPs) calculated from pedigree or "omics" data, there are numerous cases in which candidates differ in their population parameters, most notably the prediction accuracy (ρ) for the TGVs. In genomic selection, ρ strongly depends on the size of the training set and its relationship to the prediction set (e.g., Auinger et al. 2021; Clark et al. 2012; Habier et al. 2007). As demonstrated by experimental studies and simulations, adding more half-sibs to full-sibs in the training set improves ρ for genomic prediction within full-sib families (Brauner et al. 2019; Lehermeier et al. 2014; Lian et al. 2014; Riedelsheimer et al. 2013). Additionally, if pedigree, genomic, metabolic, or transcriptomic data are collected for different sets, the prediction accuracy of BLUPs calculated from different "omics" features or combinations of them can vary significantly among candidates (Seifert et al. 2018; Westhues et al. 2017; Zenke-Philippi et al. 2017). The same holds true for recently proposed approaches of phenomic selection based on sensor data and NIRS measurements (Robert et al. 2022; Weiß et al. 2022). Thus, breeders should be aware of the consequences of different prediction accuracies for the composition of the selected candidates.

A related, albeit slightly distinct scenario unfolds in hybrid breeding. Typically, lines from two genetically distant parent populations are selected based on predictors of their general combining ability (GCA) to attain a high selection response in the predicted hybrids (Melchinger et al. 2023). In general, breeders select an equal proportion of lines from each parent population for producing a factorial of hybrids among them (Melchinger and Posselt 2013). However, this approach may not be optimal if the two parent populations differ in their GCA variances and/or prediction accuracy for GCA effects. To our knowledge, no research has addressed the determination of the optimal proportion of lines to be selected from each parent population under this scenario.

The main objective of this study was to quantify and analyze the expected selection response when applying truncation selection to candidates from two sets differing in their population parameters. First, we extend Cochran's formula for determining how the selection response in the combined set and the composition of the selected fraction depends on the proportion and selection response of the individual sets. Second, we derive solutions to determine the threshold, or equivalently the selected proportion, in each set to maximize the selection response in the combined set and examine the implications for selection based on BLUPs or BLUEs. Third, we explore how to optimize the selection response in hybrid breeding if the female and male parent lines of a complete factorial are selected based on their predicted GCA and the two parent populations differ in the variance and/or prediction accuracy of GCA. We augment our theoretical findings with numerical calculations that assess the benefits of utilizing optimal selected proportions and their impact on the composition of the selected set.

Theory

The results in this section are given for two sets of genotypes Π_1 and Π_2 that can originate from the same or different populations, but they can be extended to any number of sets. The two disjoint sets Π_1 and Π_2 can be of unequal size with proportions π_1 and $\pi_2 = 1 - \pi_1$, respectively, in the combined set $\Pi_1 \cup \Pi_2$. We assume that the SC for the candidates from Π_1 or Π_2 is identically independently distributed according to normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Under these assumptions, applying truncation selection with threshold t_1 and t_2 to the candidates in Π_1 and Π_2 corresponds directly to selecting proportions $\alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)$ and $\alpha\left(\frac{t_2-\mu_2}{\sigma_2}\right)$, respectively. Here, $\alpha(x)$ denotes the proportion

selected from a normal distribution $N(0,1)$ using threshold x, and $i_{\alpha(x)}$ represents the corresponding selection intensity.

In order to simplify formulas, we will use the abbreviation $\boldsymbol{\xi} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1)$. Thus, we get for the proportion of candidates selected from $\Pi_1 \cup \Pi_2$ using thresholds t_1 and t_2 ("Appendix 1," Eq. 17)

$$\alpha_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi}) = \alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right)\pi_1 + \alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right)\pi_2 \tag{1}$$

and for the proportion of candidates from Γ_1 and Γ_2 in the selected fraction $\Gamma_1 \cup \Gamma_2$ ("Appendix 1," Eq. 18)

$$\begin{aligned}\gamma_1(t_1, t_2, \boldsymbol{\xi}) &= \frac{\alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)\pi_1}{\alpha_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi})} = \frac{|\Gamma_1|}{|\Gamma_1 \cup \Gamma_2|} \text{ and} \\ \gamma_2(t_1, t_2, \boldsymbol{\xi}) &= \frac{\alpha\left(\frac{t_2-\mu_2}{\sigma_2}\right)\pi_2}{\alpha_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi})} = \frac{|\Gamma_2|}{|\Gamma_1 \cup \Gamma_2|} = 1 - \gamma_1(t_1, t_2, \boldsymbol{\xi})\end{aligned} \tag{2}$$

Assuming the regression coefficient of the SC on the TGV is b_1 in Π_1 and b_2 in Π_2, and applying the breeders' equation for each set, we get for the total selection response of TGVs under truncation selection in $\Pi_1 \cup \Pi_2$ with thresholds t_1 and t_2 ("Appendix 1," Eq. 22)

$$\begin{aligned}&\Delta G_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi}, b_1, b_2) \\ &= \Delta G_1\big((t_1 - \mu_1), \sigma_1, b_1\big)\gamma_1(t_1, t_2, \boldsymbol{\xi}) \\ &\quad + \Delta G_2\big((t_2 - \mu_2), \sigma_2, b_2\big)\gamma_2(t_1, t_2, \boldsymbol{\xi}) \\ &\quad + \mu_1\left[\gamma_1(t_1, t_2, \boldsymbol{\xi}) - \pi_1\right] \\ &\quad + \mu_2[\gamma_2(t_1, t_2, \boldsymbol{\xi}) - \pi_2]\end{aligned} \tag{3}$$

where $\Delta G_1\big((t_1 - \mu_1), \sigma_1, b_1\big) = b_1\sigma_1 i_{\alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)}$ and $\Delta G_2\big((t_2 - \mu_2), \sigma_2, b_2\big) = b_2\sigma_2 i_{\alpha\left(\frac{t_2-\mu_2}{\sigma_2}\right)}$ refer to the selection response in set Π_1 and Π_2, respectively.

A special situation exists in hybrid breeding with two genetically distant parent populations, where Π_1 and Π_2 correspond to sets of lines from the seed or pollen parent population, respectively. The TGV refers to the general combining ability (GCA) of each line in cross-combinations with the other parent population. Since GCA values are defined as deviations from the overall mean of the hybrid population $\Pi_1 \times \Pi_2$, we assume that the SC for the GCA of the lines from Π_1 and Π_2 follows normal distributions with $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively, and the regression coefficients of the TGV of GCA effects on the SC are b_1 and b_2, respectively. In phenotypic selection, the SC is commonly based on the testcross performance of each line evaluated in crosses with one or several tester(s) from the opposite population of the heterotic pattern. In genomic selection, GCA can be predicted from the marker profile of the parent lines and phenotypic data of hybrids in a training set (cf. Bernardo 1996; Technow et al. 2014).

The lines with highest predicted GCA effects in each parent population are generally selected for producing a factorial to be phenotyped in the final step of cultivar development (Melchinger and Posselt 2013). Thus, the selection response ΔG_{Hyb} in the complete factorial $\Gamma_1 \times \Gamma_2$ of hybrids, produced by mating set Γ_1 of GCA-selected lines from Π_1 with set Γ_2 of GCA-selected lines from Π_2, compared to the factorial $\Pi_1 \times \Pi_2$ among unselected lines, is equal to the sum of the selection response for GCA effects $\Delta G_1(t_1, \sigma_1, b_1) = b_1\sigma_1 i_{\alpha\left(\frac{t_1}{\sigma_1}\right)}$ plus $\Delta G_2(t_2, \sigma_2, b_2) = b_2\sigma_2 i_{\alpha\left(\frac{t_2}{\sigma_2}\right)}$ in parent population Π_1 and Π_2, respectively, and we have for $\boldsymbol{\theta} = (\sigma_1, \sigma_2, b_1, b_2)$

$$\Delta G_{\text{Hyb}}(t_1, t_2, \boldsymbol{\theta}) = \Delta G_1(t_1, \sigma_1, b_1) + \Delta G_2(t_2, \sigma_2, b_2) \tag{4}$$

where $\alpha(\frac{t_1}{\sigma_1}) = \frac{|\Gamma_1|}{|\Pi_1|}$ and $\alpha(\frac{t_2}{\sigma_2}) = \frac{|\Gamma_2|}{|\Pi_2|}$ is the proportion of selected lines in Π_1 and Π_2, respectively. Note that

$$\alpha_{\text{Hyb}}(t_1, t_2, \boldsymbol{\theta}) = \alpha\left(\frac{t_1}{\sigma_1}\right) \times \alpha\left(\frac{t_2}{\sigma_2}\right) = \frac{|\Gamma_1|}{|\Pi_1|} \times \frac{|\Gamma_2|}{|\Pi_2|} = \frac{|\Gamma_1 \times \Gamma_2|}{|\Pi_1 \times \Pi_2|} \tag{5}$$

corresponds to the proportion of hybrids in $\Gamma_1 \times \Gamma_2$ selected in silico from the set of all possible hybrids in $\Pi_1 \times \Pi_2$.

Maximizing the total selection response by optimal choice of thresholds

Depending on the budget and size of the breeding program, the breeder has restrictions on the total number of genotypes to be selected from the candidates in a given cycle. This applies irrespective of whether the selected candidates are promoted to further testing for cultivar development or recombined to generate new base material for the next breeding cycle in recurrent selection. Therefore, the total proportion of selected candidates (α_T) is typically fixed. Nevertheless, the breeder still has the option to optimize the total selection response in $\Pi_1 \cup \Pi_2$ by selecting different proportions of candidates from Π_1 and Π_2, respectively, while keeping $\alpha_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi})$, the total proportion of genotypes selected from $\Pi_1 \cup \Pi_2$, fixed. Thus, the goal is to find thresholds t_1^* and t_2^*, or equivalently selected proportions $\alpha_1^* = \alpha\left(\frac{t_1^* - \mu_1}{\sigma_1}\right)$ and $\alpha_2^* = \alpha\left(\frac{t_2^* - \mu_2}{\sigma_2}\right)$, which maximize the total selection response $\Delta G_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi}, b_1, b_2)$ under the side condition $\alpha_{\text{Tot}}(t_1, t_2, \boldsymbol{\xi}) = \alpha_T$.

A solution to this problem can be obtained by applying a Lagrange multiplier approach. Our derivations show that (t_1^*, t_2^*) are obtained as solutions of the following equations in (t_1, t_2) ("Appendix 2," Eqs. 30 and 31):

$$t_1 = \frac{b_2(t_2-\mu_2)+\mu_2-\mu_1}{b_1}+\mu_1 \text{ or equivalently}$$
$$t_2 = \frac{b_1(t_1-\mu_1)+\mu_1-\mu_2}{b_2}+\mu_2 \quad (6)$$

and

$$\alpha_{\text{Tot}}(t_1,t_2,\boldsymbol{\xi}) = \alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)\pi_1+\alpha\left(\frac{t_2-\mu_2}{\sigma_2}\right)(1-\pi_1)=\alpha_{\text{T}}. \quad (7)$$

Solutions (t_1^*, t_2^*) of these equations can be obtained by mathematical software, such as Mathematica (Wolfram 1999), and subsequently used to calculate α_1^*, α_2^* and $\Delta G_{\text{Tot}}(t_1^*, t_2^*, \boldsymbol{\xi}, b_1, b_2)$. In order to assess the improvement in the total selection response, which can be achieved by applying optimal thresholds (t_1^*, t_2^*) instead of identical thresholds $t_1^i = t_2^i$ for both sets satisfying the side condition $\alpha_{\text{Tot}}(t_1^i, t_2^i, \boldsymbol{\xi}) = \alpha_{\text{T}}$ in Eq. 7, we suggest using the ratio

$$\Psi_{\text{Tot}}(\alpha_{\text{T}},\boldsymbol{\xi},b_1,b_2) = 100\times\left[\frac{\Delta G_{\text{Tot}}(t_1^*,t_2^*,\boldsymbol{\xi},b_1,b_2)-\Delta G_{\text{Tot}}(t_1^i,t_2^i,\boldsymbol{\xi},b_1,b_2)}{\Delta G_{\text{Tot}}(t^i,t^i,\boldsymbol{\xi},b_1,b_2)}\right]. \quad (8)$$

In hybrid breeding, the breeder is also limited in terms of the number of promising predicted hybrids that can be evaluated in a factorial for product development in the next step of the breeding scheme. Thus, the goal is to find optimal proportions α_1^o and α_2^o of candidates from Π_1 and Π_2, or equivalently optimal thresholds t_1^o and t_2^o, for selection in Π_1 and Π_2, respectively, which maximize the selection response $\Delta G_{\text{Hyb}}(t_1, t_2, \boldsymbol{\theta})$ for the factorial produced between the GCA-selected lines. However, instead of using Eq. 7, the side condition takes the form

$$\alpha_{\text{Hyb}}(t_1,t_2,\boldsymbol{\theta}) = \alpha\left(\frac{t_1}{\sigma_1}\right)\times\alpha\left(\frac{t_2}{\sigma_2}\right)=\alpha_{\text{H}}, \quad (9)$$

where α_{H} is the fixed proportion of hybrids to be selected for testing in the final stage of hybrid development.

A solution to this maximization problem can be found again by applying a Lagrange multiplier approach ("Appendix 3"). Accordingly, thresholds (t_1^o, t_2^o) optimizing $\Delta G_{\text{Hyb}}(t_1, t_2, \boldsymbol{\theta})$ in the factorial of hybrids among selected lines are found as solutions (t_1, t_2) ("Appendix 3," Eqs. 43) of Eq. 9 and

$$b_2t_2 - b_1t_1 + \Delta G_1(t_1,\sigma_1,b_1) - \Delta G_2(t_2,\sigma_2,b_2) = 0. \quad (10)$$

Numerical solutions for (t_1^o, t_2^o) can be obtained by mathematical software such as Mathematica and subsequently used to calculate the proportions $\alpha_1^o = \alpha\left(\frac{t_1^o}{\sigma_1}\right)$ and $\alpha_2^o = \alpha\left(\frac{t_2^o}{\sigma_2}\right) = \alpha_{\text{H}}/\alpha_1^o$ to be selected in Π_1 and Π_2, respectively, and finally, $\Delta G_{\text{Hyb}}(t_1^o, t_2^o, \boldsymbol{\theta})$.

In order to assess the improvement in the total selection response, which can be achieved by using the optimal proportions (α_1^o, α_2^o) compared to selecting an equal proportion $\alpha^e = \sqrt{\alpha_{\text{H}}}$ of lines from each population, i.e., using thresholds $t_1^e = \sigma_1\Phi^{-1}(1-\alpha^e)$ and $t_2^e = \sigma_2\Phi^{-1}(1-\alpha^e)$, we suggest using the ratio

$$\Psi_{\text{Hyb}}(\alpha_{\text{H}},\boldsymbol{\theta}) = 100\times\left[\frac{\Delta G_{\text{Hyb}}(t_1^o,t_2^o,\boldsymbol{\theta})-\Delta G_{\text{Hyb}}(t_1^e,t_2^e,\boldsymbol{\theta})}{\Delta G_{\text{Hyb}}(t_1^e,t_2^e,\boldsymbol{\theta})}\right]. \quad (11)$$

Application to selection based on BLUPs

Let u denote the random variable of true breeding values (TBVs) and $\hat{u}$ their BLUPs, obtained by the use of pedigree or "omics" data. As shown by Henderson (1975), the standard deviation σ_u of TGVs and the standard deviation σ of their BLUPs are related by $\sigma = \rho\sigma_u$, where ρ is the prediction accuracy, reflecting the shrinkage of BLUPs compared to the TBVs. Hence, we have $\sigma_1 = \rho_1\sigma_{u_1}$ and $\sigma_2 = \rho_2\sigma_{u_2}$. Further, the regression of u on $\hat{u}$ is equal to 1.0 for each set, so that $b_1 = 1.0$ and $b_2 = 1.0$ and this result holds true under fairly general conditions ("Appendix 4"). Thus, from Eq. 6 we obtain $t_1^* = t_2^*$, even if $\mu_1 \neq \mu_2$, $\sigma_{u_1}^2 \neq \sigma_{u_2}^2$, and $\rho_1 \neq \rho_2$. Consequently, using identical thresholds for the predicted values of TGVs (calculated as BLUPs plus the mean μ of the corresponding set) maximizes the selection response in the combined set. In conclusion, for BLUPs there is no need to search for the optimal threshold in each set and one must merely find the common threshold $t^* = t_1^* = t_2^*$ for both sets satisfying the side condition in Eq. 7, which can be obtained by solving the equation

$$\Phi\left(\frac{t^*-\mu_1}{\sigma_1}\right)\pi_1+\Phi\left(\frac{t^*-\mu_2}{\sigma_2}\right)(1-\pi_1)=1-\alpha_{\text{T}} \quad (12)$$

Moreover, the total selection response in the combined set $\Pi_1 \cup \Pi_2$ for the common threshold t^* is

$$\begin{aligned}&\Delta G_{\text{Tot-BLUP}}(t^*,t^*\boldsymbol{\xi},1,1)\\ &=\frac{1}{\alpha_{\text{T}}}[\sigma_1\varphi\left(\frac{t^*-\mu_1}{\sigma_1}\right)\pi_1+\sigma_2\varphi\left(\frac{t^*-\mu_2}{\sigma_2}\right)\pi_2\\ &+\mu_1\pi_1\left(\alpha\left(\frac{t^*-\mu_1}{\sigma_1}\right)-\alpha_{\text{T}}\right)+\mu_2\pi_2\left(\alpha\left(\frac{t^*-\mu_2}{\sigma_2}\right)-\alpha_{\text{T}}\right)]\end{aligned} \quad (13)$$

Application to selection based on BLUEs

In phenotypic selection (PS) based on BLUEs, the regression of TGVs on the SC is equal to their heritability (Falconer and Mackay 1996, p. 189), so that $b_1 = h_1^2$ and $b_2 = h_2^2$. Further, the standard deviation σ_u of TBVs and the standard deviation σ of their BLUEs used in PS are related by $\sigma = \frac{\sigma_u}{h}$.

Hence, we have $\sigma_1 = \sigma_{u_1}/h_1$ and $\sigma_2 = \sigma_{u_2}/h_2$. Thus, Eq. 3 becomes

$$\begin{aligned}&\Delta G_{\text{Tot - PS}}\left(t_1, t_2, \xi, h_1^2, h_2^2\right)\\&= h_1^2\sigma_1\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\pi_1 + h_2^2\sigma_2\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right)\pi_2\\&\quad+\mu_1\left[\gamma_1\left(t_1, t_2, \xi\right)-\pi_1\right]+\mu_2\left[\gamma_2\left(t_1, t_2, \xi\right)-\pi_2\right]\end{aligned} \tag{14}$$

From Eqs. 6 and 7, the optimal choice of thresholds t_1^* and t_2^* are obtained as solutions of

$$\begin{aligned}&\Phi\left(\frac{t_1^*-\mu}{\sigma_1}\right)\pi_1 + \Phi\left(\frac{h_1^2\left(t_1^*-\mu_1\right)+\mu_1-\mu_2}{h_2^2\sigma_2}\right)\pi_2 = 1-\alpha_{\text{T}} \text{ and}\\&t_2^* = \frac{h_1^2\left(t_1^*-\mu_1\right)+\mu_1-\mu_2}{h_2^2}+\mu_2\end{aligned} \tag{15}$$

Numerical analyses

All equations in the theory part were programmed in software Mathematica (Wolfram 1999) for numerical analyses. As a first check for Eq. 6 and the derivations in "Appendix 2," we numerically compared the selection response ΔG_{Tot} for BLUPs achieved with optimized thresholds (t_1^*, t_2^*) versus identical ($t_1^i = t_2^i$) thresholds for BLUPs, setting $b_1 = b_2 = 1.0$ in our program. Regardless of the means (μ_1, μ_2) and standard deviations (σ_1, σ_2) of the SC in Π_1 and Π_2, as well as the choice of π_1 and α_{T}, the value of ΔG_{Tot} obtained for (t_1^*, t_2^*) and ($t_1^i = t_2^i$) were identical except for tiny differences attributable to rounding errors so that Ψ_{Tot} was practically zero (data not shown), confirming our theoretical results.

For BLUEs, we calculated on one hand the values for Ψ_{Tot} and $\gamma_1^* = \gamma_1\left(t_1^*, t_2^*, \xi\right)$ obtained by using the solutions for (t_1^*, t_2^*) obtained with the Lagrange multiplier approach (Eq. 6). On the other hand, we used Function NMaximize in Mathematica to determine the maximum of ΔG_{Tot} under the side condition in Eq. 7. Again, the numerical results from both calculations were in perfect agreement except for numerical inaccuracies.

For finding the maximum selection response ΔG_{Hyb} in the hybrid population $\Pi_1 \times \Pi_2$, we used function NMaximize in Mathematica in combination with the side condition in Eq. 9 to find the optimum choice of selected proportions(α_1^o, α_2^o). These values we used to calculate according to Eq. 8 the percentage improvement (ψ_{Hyb}) in ΔG_{Hyb} when using optimized (α_1^o, α_2^o) instead of equal $\left(\alpha_1^e = \alpha_2^e\right)$ proportions of lines selected from population Π_1 andΠ_2.

For investigating the consequences of BLUE-based selection on the magnitude ofΨ_{Tot}, γ_1^* and $\gamma_1^i = \gamma_1\left(t_1^i, t_2^i, \xi\right)$ as a function of other relevant population parameters, we made the assumption without loss of generality that $\mu_1 = 0$ and $\sigma_{u_1} = 1.0$. This can be achieved by centering the original SC values of all candidates as deviations from μ_1 and dividing them by $\sigma_1 = \sigma_{u_1}/h_1$. Moreover, for representing Ψ_{Tot} and γ_1^* or γ_1^i in contour plots as functions of μ_2 andh_2, we assumed $h_1^2 = \sqrt{0.5}$ and identical genetic standard deviations in both sets ($\sigma_{u_1} = \sigma_{u_2}$), which closely approximates the conditions encountered in many situations in plant breeding programs.

Software availability statement

The Mathematica programs developed for the numerical analyses of this study are available at https://github.com/TUMplantbreeding/AEM/Opt_selection_with_multiple_sets and can be downloaded from there.

Results

Figure 1 examines for BLUPs the shift in the proportion of candidates from Π_1 before (π_1) and after selection (γ_1^*). We present the ratio $\gamma_1^* : \pi_1$ as a function of $\mu_2$andρ_2 under the assumptions mentioned above (μ_1 =0,$\sigma_{u_1}^2 = \sigma_{u_2}^2 = 1.0$, $\rho_1 = 0.50$). Regardless of the magnitude of α_{T} andπ_1, the contour lines were straight lines, indicating that γ_1^* depends on a linear function of μ_2 and ρ_2 with weights of these parameters determining their slope. For small values of π_1 or α_{T}, the ratio reduced substantially with an increasing sum $\mu_2 + \rho_2$ so that even for moderate values for one of these parameters, the ratio was smaller than 0.1, indicating that less than 10% of the initial proportion π_1 was recovered in γ_1^*. For π_1= 0.90 in combination $\alpha_{\text{T}} \geq 0.10$, the ratio was less affected by increasing ρ_2 and reduced only moderately with increasing μ_2, yet the slope of the contour lines changed with μ_2.

As expected, under optimal thresholds (t_1^*, t_2^*) for selection based on BLUEs, the contour plots for $\gamma_1^* : \pi_1$ were identical to those obtained for BLUPs, when replacing ρ_1 by $h_1 = \sqrt{h_1^2}$ and ρ_2 by$h_2 = \sqrt{h_2^2}$, respectively (results not shown). For comparison, we also analyzed the ratio $\gamma_1^i : \pi_1$ as a function of μ_2 and h_2 to monitor the relative change in the proportion of candidates from Π_1 before (π_1) and after selection (γ_1^i) based on BLUEs with identical thresholds ($t_1^i = t_2^i$) for both sets (Supplementary Figure 1). Compared with $\gamma_1^* : \pi_1$, the ratio $\gamma_1^i : \pi_1$ changed less with increasing μ_2 and h_2, particularly for large values of π_1 or α_{T}. The ratio depended mainly on the magnitude of h_2 and less on the size of μ_2. For $\pi_1 \leq 0.50$ and $\alpha_{\text{T}} \geq 0.10$, the ratio was smaller or larger than 1.0 if h_2 falls below or exceeds h_1,

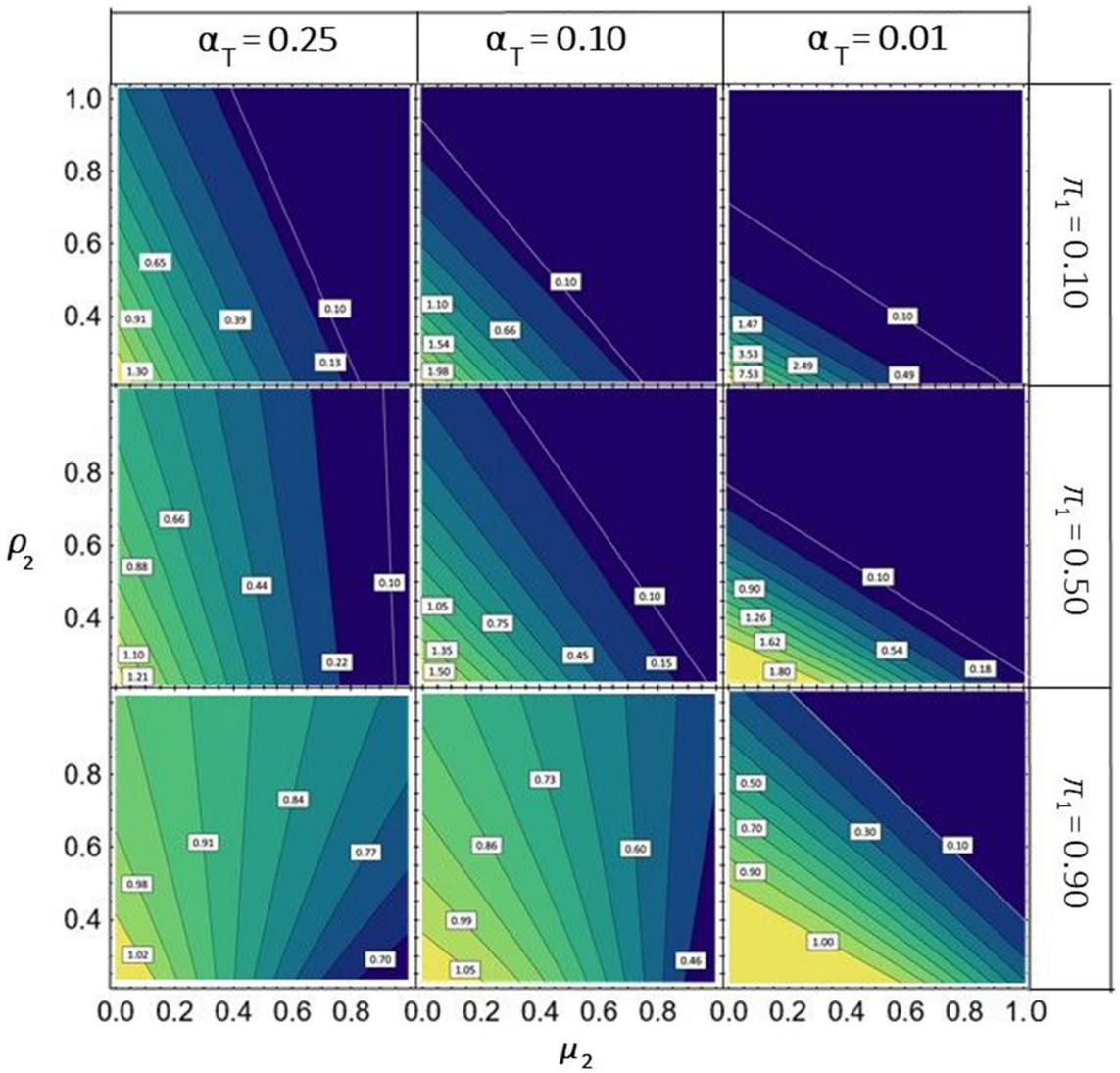

Fig. 1 Contour plots for the ratio$\gamma_1^* : \pi_1$, indicating the shift in the proportion of genotypes from Π_1 before (π_1) and after (γ_1^*) truncation selection based on BLUPs, when using optimal (=identical) thresholds ($t_1^* = t_2^*$) in set Π_1 andΠ_2. The graphs show $\gamma_1^* : \pi_1$ as a function of the mean μ_2 and the prediction accuracy ρ_2 of the selection criterion (SC) in Π_2 for various values of π_1 and α_T, the proportion of candidates selected from $\Pi_1 \cup \Pi_2$. Assumptions are μ_1=0, $\sigma^2_{u_1} = \sigma^2_{u_2} = 1.0, \rho_1 = 0.50$, i.e., $\boldsymbol{\xi}$ =$(0, \mu_2, 0.5, \rho_2, \pi_1)$. The white labels attached to the contour lines show the corresponding numerical values

respectively, and increasing μ_2 had only a moderately reducing effect.

When performing mild selection ($\alpha_T = 0.25$) with BLUEs, the size of Ψ_{Tot}, reflecting the improvement in overall selection response achieved by using optimal (t_1^*, t_2^*) instead of identical thresholds($t_1^i = t_2^i$), was consistently smaller than 10%, irrespective of π_1 and the investigated range of μ_2 and h_2 (Fig. 2). For $\alpha_T = 0.10$, Ψ_{Tot} was close to zero for π_1=0.1 but exceeded 10% for π_1=0.50 and high values of h_2. Under stringent selection with $\alpha_T = 0.01$ and $h_2 \geq 0.90$, Ψ_{Tot} surpassed 20% for π_1=0.50, regardless of μ_2, or if $\pi_1 = 0.50$ and $h_2 \geq 0.5$. Setting μ_2= 1.0 had only a minor effect on increasing Ψ_{Tot} compared to increasing h_2 from $\sqrt{0.5}$ to 0.9.

Figure 3 shows Ψ_{Hyb}, the increase in selection response for hybrids when selecting optimal (α_1^o, α_2^o) versus equal ($\alpha_1 = \alpha_2 = \alpha^e = \sqrt{\alpha_H}$) proportions of lines from each parent population, as a function of $\sigma_2 : \sigma_1$, the ratio of the standard deviations of BLUPs for GCA effects of lines in Π_1 and Π_2. Ψ_{Hyb} showed an approximately quadratic decrease with increasing the ratio $\sigma_2 : \sigma_1$ from 0.5 to 1.0 and minor differences for different values of α_H. For $\sigma_2 : \sigma_1 = 0.5$, Ψ_{Hyb} was approximately 6% for all values of α_H. The ratio α_1^o: α^e displayed a quadratic decrease with increasing $\sigma_2 : \sigma_1$ with large differences depending on α_H. For $\sigma_2 : \sigma_1 = 0.5$ and $\alpha_H \leq 0.01$, α_1^o: α^e was smaller than 0.25, reflecting that selection of hybrids relied almost entirely on stringent GCA

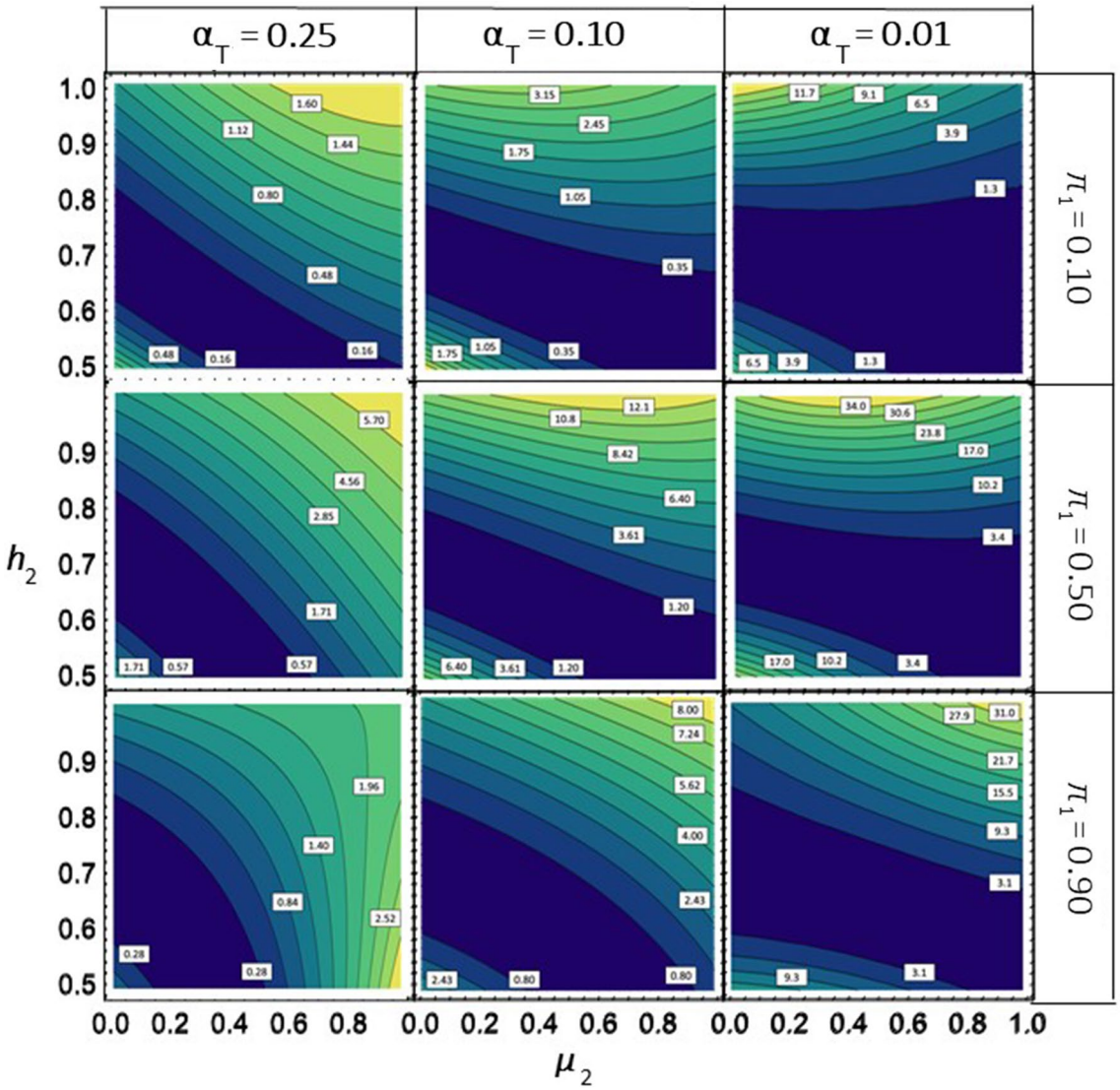

Fig. 2 Contour plots for $\Psi_{\mathrm{Tot}}(\alpha_T, \xi, h_1^2, h_2^2)$, indicating the percentage increase of the selection response ΔG_{Tot} for selection based on BLUEs in $\Pi_1 \cup \Pi_2$, when using optimal (t_1^*, t_2^*) versus identical $(t_1^i = t_2^i)$ thresholds for truncation selection in set Π_1 andΠ_2, respectively. The graphs show Ψ_{Tot} as a function of the mean μ_2 and h_2, the square root of the heritability of the BLUEs in Π_2 for various values of π_1 and α_T, the proportion of candidates selected from $\Pi_1 \cup \Pi_2$. Assumptions are $\mu_1 = 0$, $\sigma^2_{u_1} = \sigma^2_{u_2} = 1.0, h_1^2 = 0.50$, i.e., $\xi = (0, \mu_2, \sqrt{2}, \frac{1}{\sqrt{h_2^2}}, \pi_1)$. The white labels attached to the contour lines show the corresponding numerical values

selection of lines in the parent population with higher variance of BLUPs and only mild selection in the other parent population.

Discussion

Examples of sets differing in population parameters

In all breeding categories described by Schnell (1982), plant breeders generally evaluate and select genotypes from multiple families in parallel as evident from publications on public and private breeding programs in maize and wheat (e.g., Auinger et al. 2021; Bonnett et al. 2022; Lian et al. 2014). The parents of these mostly bi-parental families generally differ in their performance level and relationship, and therefore, the progenies differ with respect to relevant population parameters. Nevertheless, these materials are routinely evaluated together in the same experiment(s) and genotypes promoted to the next stage of the program are often selected without giving much attention to their origin.

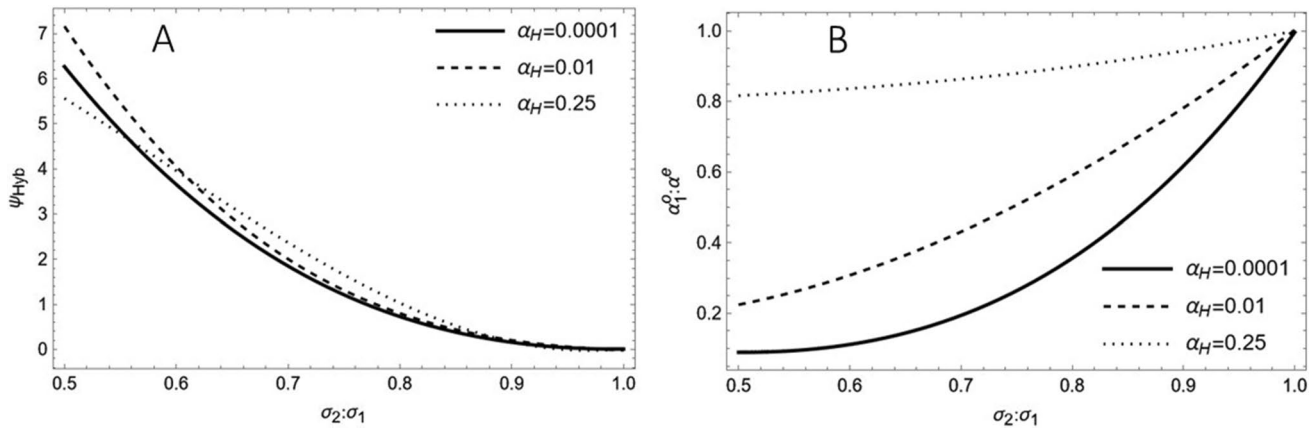

Fig. 3 **A** Percentage increase $\psi_{\mathrm{Hyb}}(\alpha_{\mathrm{H}}, \theta)$ of the selection response in the hybrid population $\Pi_1 \times \Pi_2$ and **B** ratio of the optimal proportion of selected candidates (α_1^0) from Π_1 versus an equal ($\alpha^e = \sqrt{\alpha_{\mathrm{H}}}$) proportion of lines selected from each parent population based on GCA predicted by BLUPs for $\theta = (\sigma_1, \sigma_2, 1, 1)$. The graphs show ψ_{Hyb} and as function of $\sigma_2 : \sigma_1$, the ratio of standard deviations of BLUPs for GCA of lines in Π_2 and Π_1, respectively, for different values of α_{H}, the proportion of hybrids selected from $\Pi_1 \times \Pi_2$

A comparable situation exists in introgression breeding programs when multiple populations are developed by crossing elite germplasm with various donors (e.g., Barbosa et al. 2021). These materials generally differ in their performance level and genetic variance due to disparate adaptation of the donors to the target environment(s) and varying proportions of donor germplasm in the pedigree. In pre-breeding programs too, the differences among populations can be extremely large as reported for landraces of maize (Böhm et al. 2017; Hölker et al. 2019; Mayer et al. 2017). If all populations are evaluated in a common experiment, breeders are inclined to apply the same threshold for identifying superior candidates used for further breeding.

Even when dealing with a single population, so that the mean and genetic variance are identical, sets of genotypes often differ with regard to the prediction accuracy of the SC for the TGV of candidates. This can be attributable to unbalanced data from multi-environment trials, where some sets are evaluated in fewer environments or replications than others. For instance, top performers remain in the testing pipeline for several years, while new entries are added to the system (Piepho et al. 2008). Moreover, some genotypes might be tested less intensively owing to problems in seed multiplication, as occurs in the production of doubled-haploid lines (Chaikam et al. 2019) or in speed breeding programs (Watson et al. 2018). Further, when complex traits are monitored using sensor-based techniques (NIRS, optical sensors, etc.) or "omics" data (genomic, phenomic, etc.), the prediction accuracy tends to be notably higher in the calibration set compared to the prediction set (Melchinger and Frisch 2023) and in sets combining different "omics" features (Schrag et al. 2018; Westhues et al. 2019). Thus, there are numerous scenarios where sets of germplasm in a breeding program differ in their population parameters and breeders should be prepared to deal adequately with these situations and be aware of the implications for selection.

Contrasting BLUEs and BLUPs as selection criteria

Until two decades ago, selection decisions in plant breeding relied exclusively on BLUEs of the candidates, a practice that still endures in many smaller breeding programs today. Two major reasons contribute to this conservative attitude. Firstly, for traits with high heritability on an entry-mean basis, the ranking of candidates based on BLUEs and BLUPs is mostly similar. Secondly, calculation of BLUEs is straightforward and does not require information on the relationships among candidates or estimates of genetic variance components, which are challenging to obtain due to the small size of sets and rapid change over selection cycles.

Building upon the pioneering research of Henderson (1975) and inspired by the tremendous progress in animal breeding subsequent to the adoption of BLUPs, Bernardo (1994) spearheaded the implementation of BLUPs into plant breeding. With balanced data and when candidates are unrelated or possess identical co-ancestries so that their TGVs are predicted with equal accuracy, the ranking of candidates based on BLUEs and BLUPs is identical (Kennedy and Sorenson 1988). Otherwise, BLUPs offer a notable advantage by capitalizing on information from relatives and/or accommodating an efficient analysis of unbalanced data (Bernardo 2002; Piepho et al. 2008).

Another major advantage of BLUPs over BLUEs is their ability to allow direct comparisons across different breeding sets, regardless of their origin. As outlined in Eq. 6, applying the same selection threshold to the BLUPs of all

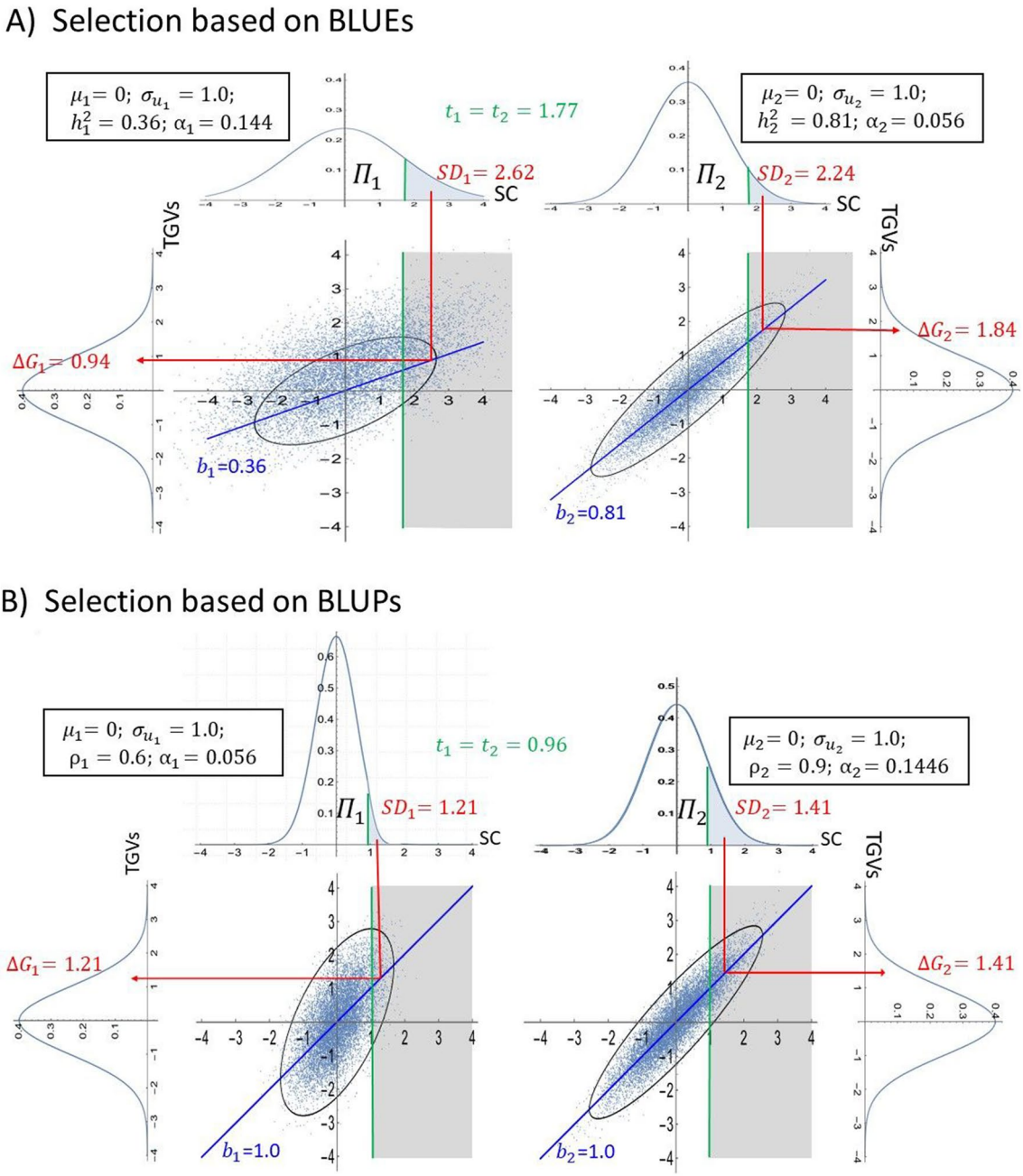

Fig. 4 Individual and joint probability density functions (pdf) of the true genetic values (TGV ~ $N(0,1)$) and selection criterion (SC ~ $N(0,\sigma^2)$ for sets Π_1 and Π_2 with equal proportions ($\pi_1 = \pi_2 = 0.5$). **A** SC are BLUEs with $\sqrt{h_1^2} = 0.6$ and $\sqrt{h_2^2} = 0.9$ in Π_1 and Π_2, respectively. **B** SC are BLUPs with $\rho_1 = 0.6$ and $\rho_2 = 0.9$ in Π_1 and Π_2, respectively. In both cases, truncation selection with identical thresholds is practiced in $\Pi_1 \cup \Pi_2$ to achieve $\alpha_T = 0.1$. SD refers to the selection differential

candidates is optimal, whereas for BLUEs distinct thresholds must generally be found to maximize the selection response of the entire program. Following Cochran (1951), our theoretical results were derived assuming that the SC and TGVs are independently and identically distributed within each set because otherwise the already complex algebra would become even more unwieldy. This assumes an idealized situation, which is seldom met in practice as data are generally unbalanced and candidates commonly differ in their relationships. However, considering that the regression function of TGVs on BLUPs remains an identity matrix even under less stringent assumptions

("Appendix 4"), we conjecture that our results for BLUPs hold approximately true across a broad spectrum of scenarios, but this warrants further research.

The difference between BLUEs and BLUPs is illustrated by two sets Π_1 and Π_2 with equal proportion ($\pi_1 = 0.5$) of unrelated candidates sampled from the same population and selection of $\alpha_T = 0.10$ candidates across $\Pi_1 \cup \Pi_2$ (Fig. 4). Thus, the two sets share identical means ($\mu_1 = \mu_2 = 0$) and genetic standard deviations ($\sigma_{u_1} = \sigma_{u_2} = 1$). Regarding the prediction accuracy of the SC, we assume $\sqrt{h_1^2} = \rho_1 = 0.6$ for Π_1 and $\sqrt{h_2^2} = \rho_2 = 0.9$ for Π_2, i.e., these values differ between the two sets but are identical for BLUEs and BLUPs within each set.

When using BLUEs, the standard deviation of the SC is larger in Π_1 compared to Π_2 ($\sigma_1 = \frac{\sigma_{u_1}}{h_1} = 1.67$ vs. $\sigma_2 = \frac{\sigma_{u_2}}{h_2} = 1.11$) due to the lower heritability. Utilizing identical thresholds ($t_1^i = t_2^i = 1.77$) for $\alpha_T = 0.10$, a larger proportion of candidates is selected in Π_1 than in Π_2 ($\alpha_1 = 0.14$ vs. $\alpha_2 = 0.06$), leading to lower selection intensity in Π_1 ($i_{\alpha_1} = 1.57$ vs. $i_{\alpha_2} = 2.02$). While the selection differentials are similar ($SD_1 = 2.62$ vs. $SD_2 = 2.24$), the selection response almost doubles in Π_2 compared to Π_1 ($\Delta G_1 = 0.94$ vs. $\Delta G_2 = 1.82$) owing to the higher heritability. Since the proportion of candidates selected from Π_1 is much larger than it would be with optimal thresholds ($\gamma_1^i = 0.72$ vs. $\gamma_1^* = 0.28$), this explains why for BLUEs the selection response $\Delta G_{\text{Tot}}(1.77, 1.77, \xi, 0.36, 0.81) = 1.19$ is significantly smaller than the maximum selection response $\Delta G_{\text{Tot}}(2.65, 1.18, \xi, 0.36, 0.81) = 1.36$ achieved with optimal thresholds $(t_1^* = 2.65, t_2^* = 1.18)$, resulting in $\psi_{\text{Tot}} = 14.5\%$. For very stringent selection with $\alpha_T = 0.01$, we get $\gamma_1^i = 0.95$. vs. $\gamma_1^* = 0.05$, leading to $\psi_{\text{Tot}} = 42.3\%$.

When using BLUPs as SC, candidates of Π_1 exhibit a smaller standard deviation than those of Π_2 ($\sigma_1 = \rho_1\sigma_{u_1} = 0.6$ vs. $\sigma_2 = \rho_2\sigma_{u_2} = 0.9$) due to increased shrinkage. Consequently, applying identical thresholds ($t_1^* = t_2^* = 0.96$) to both sets for achieving $\alpha_T = 0.10$ leads to a smaller proportion of candidates ($\alpha_1 = 0.06$ vs. $\alpha_2 = 0.14$) and a higher selection intensity ($i_{\alpha_1} = 2.02$ vs. $i_{\alpha_2} = 1.57$) for Π_1 compared to Π_2. Given that the regression for TGVs on BLUPs is equal to 1.0 (Henderson 1975), we obtain $\Delta G_1 = 1.21$ and $\Delta G_2 = 1.42$. Referring to Eqs. 2 and 13, we get $\gamma_1^* = 0.28$ and $\Delta G_{\text{Tot}}(0.96, 0.96, \xi, 1, 1) = 1.36$. While this example was chosen for simplicity, it underscores the fundamental disparities between BLUEs and BLUPs for selection in scenarios involving multiple sets.

Properties of BLUPs for selection

BLUPs possess several optimality properties for prediction of random effects in mixed linear models (Fernando and Gianola 1986; Henderson 1990). They have minimum prediction error variance and maximize the correlation to the TGVs in the class of linear unbiased predictors. Furthermore, when random effects adhere to a normal distribution and fixed effects in the mixed model are known, BLUPs have smallest mean-squared error among all possible predictors. Concerning truncation selection, we provided a proof in "Appendix 2" that when dealing with two sets characterized by distinct population parameters (e.g., means, variances and prediction accuracies of TGVs), utilizing a uniform threshold for the BLUPs across all candidates maximizes the selection response.

We derived this property of BLUPs through a Lagrange multiplier approach, which requires quite restrictive assumptions on the random effects u in the different sets. It is closely related to a more general selection principle (Fernando and Gianola 1986; Goffinet 1983). Accordingly, if n candidates are available and $k < n$ of them are to be chosen, then selecting the k candidates with highest conditional mean for an unobservable random variable u maximizes the expected value of the mean of u for the selected candidates. This result holds true independent on the joint distribution of the unobservable random variable u and the data. Under normality, the BLUP of u can be thought of as its conditional mean. Thus, even when the candidates are from different sets, selecting the k candidates with the highest values for BLUPs ($\hat{u}$) would maximize the response to selection and no further corrections are needed.

In a strict sense, selecting a fixed number k or constant proportion $\alpha = k/n$ of candidates from a finite population of size n differs from truncation selection. In truncation selection, the threshold is set so that the expected proportion of candidates is equal to α in a population of infinite size. When applying this fixed threshold to a sample of size n, the number of selected candidates may deviate from k. However, as the sample size increases, selecting a fixed number or proportion of candidates becomes equivalent to truncation selection. Therefore, the results derived for truncation selection in this study closely approximate those for selecting a constant proportion of candidates.

Our approach for proving the optimality property of BLUPs under truncation selection allows calculating the optimal proportion α_1^* and α_2^* of candidates selected from set Π_1 andΠ_2, respectively, given reliable estimates of the population parameters are available. This information is important for optimizing the allocation of resources in genomic selection based on BLUPs. By knowing α_1^* and α_2^* in advance, we can calculate the selection response across both the training and prediction set. Thus, we can find the

ideal balance between (1) the expenditures allocated to the training set, which determines mainly the prediction accuracies of both the training and prediction set, and (2) the size of both sets, which determines α_T. A thorough examination of this complex problem is beyond the scope of this study and warrants further research.

Using the same threshold for BLUPs does not necessarily mean that all candidates share an equal likelihood of being selected, even if they possess the same TGV as highlighted in the literature (Woolliams et al. 2015). This can be exemplified by Fig. 4, where for $\alpha_T = 0.10$ the proportion of candidates from set Π_1 would reduce from $\pi_1 = 0.50$ before selection to $\gamma_1^* = \gamma_1^i = 0.28$ after selection owing to the lower prediction accuracy for Π_1 and increased shrinkage of BLUPs.

Composition of the selected fraction

Generalizing Cochran's formula for selection response to the case of multiple sets allowed us to examine the proportions (γ_1, γ_2) of selected candidates originating from Π_1 andΠ_2. This is of interest for two reasons. First the selection response for the combined set depends on a weighted summation of the selection response in each set (Eq. 3), with weights corresponding to the post-selection fractions γ_1 and γ_2. Second, the makeup of the selected fraction is critical for further breeding progress, given that these candidates are used either directly for product development and/or for generating the base materials of the next breeding cycle. In extreme cases, ΔG_{Tot} for BLUEs can even be negative. For instance, if only mild selection ($\alpha_1 = 0.45$) is applied to the inferior, smaller population Π_1(π_1=0.2, $\mu_1 = 0$, $\sigma_1 = 1$, h_1^2= 0.36) but stringent selection ($\alpha_2 = 0.0125$) is applied to Π_2 ($\mu_2 = 2.0$, $\sigma_2 = 2$ and h_2^2= 0.81) so that $\gamma_1 = 0.90$ is much larger than π_1, the outcome would be$\Delta G_{Tot} = -0.70$.

Here, we focus our discussion on the composition of the selected fraction obtained through the use of BLUPs with a uniform threshold for all candidates. As indicated by the graphs in Fig. 1, the change in the composition of the candidates before and after selection, expressed by the ratio γ_1^*: π_1, can be striking. For instance, when $\pi_1 = 0.10$ and/or $\alpha_T = 0.01$, the proportion retained from the inferior set Π_1 dwindles to less than 10% of its original proportion, if μ_2 surpasses μ_1 by about one genetic standard deviation under otherwise identical conditions. Consequently, if materials from introgression programs are evaluated together with elite germplasm and the same threshold is applied to the BLUPs of both groups, hardly any novel germplasm will be selected due to its low performance level. Thus, it would be prudent to apply different thresholds for both groups to have a realistic chance that some of the promising new genotypes are retained for further breeding.

Likewise, the ratioγ_1^*: π_1 falls below 0.20, if two sets share equal size and population parameters, yet $\rho_2 \geq 0.68$ while $\rho_1 = 0.50$. Differences of this magnitude have been observed in genomic prediction of maize hybrids, in which case the prediction accuracy significantly decreased from H2 hybrids, where both parents are used as parents of a hybrid in the training set, to H1 and H0 hybrids, where only one or none of the parent lines, respectively, contribute to a hybrid in the training set (Seye et al. 2020; Technow et al. 2014; Westhues et al. 2017). While it seems rewarding to have a much larger number of H0 hybrids than H1 and H2 hybrids due to their lower costs (involving only production and genotyping of parent lines), the contribution of H0 hybrids to the overall selection response is generally overrated because their selected proportion is much smaller than for H2 and H1 hybrids owing to the lower prediction accuracy. Consequently, H0 hybrids contribute significantly less to the selection response than expected based on their proportion in the entire set of predictable hybrids. This aspect is crucial when optimizing the distribution of resources allocated to the training and prediction sets (Riedelsheimer and Melchinger 2013).

There are many further examples, where sets differ in their prediction accuracy because they differ in the number of close relatives in the training set. For this reason, genotypes in the training set have generally a significantly higher prediction accuracy than those in the prediction set, leading to a notable underrepresentation of the latter in the selected set. Likewise, in recycling breeding breeders typically generate more and larger families from crosses of elite parents. If the training set is sampled proportional to the size of these families, it follows that genotypes descending from the top parents have higher prediction accuracy due to more and closer relatives in the training set than genotypes descending from less prominent parents. Thus, on top of the expected high TGVs of these progenies, the smaller shrinkage of their BLUPs further increases the likelihood that they are selected. However, this carries a high risk of selecting closely related genotypes descending from a small number of top ancestors, thereby diminishing the effective population size and long-term progress in genomic selection, particularly when applying rigorous selection pressure enabled by the low costs for genotyping with modern methods (Rasheed et al. 2017). While our focus has been primarily on diverse prediction accuracies, our conclusions can be extended to scenarios where sets differ in genetic variances.

Optimal selection of parent lines in hybrid breeding

In hybrid breeding, breeders typically work with a comparable number of lines from each parent population and select, based on GCA predicted by BLUPs, a proportional number of candidates from both groups for the final testing phase in

product development (Melchinger and Frisch 2023). Figure 3 shows that this approach is optimal when the parent populations exhibit similar variances for the SC, but this is not always the case in practice. In European maize for example, GCA variance for grain yield was approximately twice as large for dent lines compared to flint lines (Schrag et al. 2006). Similarly in hybrid rye, Wilde et al. (2003) found that GCA variance for grain yield among female lines from the Petkus pool was almost four times greater than observed among male lines from the Carsten pool. Additionally, the accuracy of predicted GCA effects can differ between the parent populations due to differences in the size and intensity of phenotyping of the training set and the use of different types of testers. Furthermore, in species like rye, where the implementation of CMS for testcross seed production differs significantly between the seed and pollen parent pools, the pedigree relationship between candidates in the prediction and training set can diverge (Wilde and Miedaner 2021).

Under these scenarios, a notable enhancement ψ_{Hyb} in selection response for predicted hybrids, compared to selecting equal proportions in each population, can be achieved by opting for more stringent selection within the parent population exhibiting the larger GCA variance. The magnitude of ψ_{Hyb} depends strongly on the ratio of GCA variances in the two parent populations but showed similar curves independent of the selected proportions α_{H} (Fig. 3). Under mild selection ($\alpha_{\text{H}} = 0.25$), the optimal α-values for the two parent populations hardly differ from each other, but for stringent selection ($\alpha_{\text{H}} = 0.0001$), a much more stringent selection must be practiced in the parent population with larger GCA than smaller GCA variance, as reflected by the low ratio α_1^0 :α^e, where $\alpha^e = \sqrt{\alpha_{\text{H}}}$. As an alternative to selecting parent lines based on their predicted GCA for producing a complete factorial of hybrids, one could directly select the most promising hybrids based on the sum of the GCA of their parents. This would result in selecting a partial factorial having the form of a triangle, with the top parents being involved in more crosses than parents with lower rank and automatically takes care of differences in the GCA variance of BLUPs for each parent population. A comparison of these two selection schemes would be highly interesting for hybrid breeding but is beyond the scope of this study.

Conclusions

When practicing truncation selection with candidates from multiple sets, new aspects must be taken into consideration as compared to selection in a single homogeneous population. This is because selection progress in the entire breeding program depends not only on the selection response in each set but also on the composition of the selected fraction. A major question is how to choose the thresholds for candidates from the various sets for maximizing the selection response of the entire breeding program. In addition to the numerous advantages of BLUPs compared to BLUEs, they have the highly desirable property that a uniform threshold can be applied to all candidates for maximizing the selection response and no further adjustment for differences in the reliability of the predictors is necessary. This applies even if the sets differ in the population parameters and/or if BLUPs of different candidates are calculated from different types or combinations of "omics" data and simplifies selection decisions. However, calculation of BLUPs requires reliable estimates of the genetic variance, which is a challenge with the small sample sizes of families used in plant breeding, but this problem has been mitigated with the use of Bayesian methods (Sorenson and Gianola 2004).

Since variation in the prediction accuracy can have a strong impact on the outcome of the selected fraction and strongly reduces the effective population size under the stringent selection, we recommend to accompany genomic selection based on BLUPs with monitoring the genetic diversity of the selected candidates. Ideally, genomic selection could be combined with optimum contribution selection (Daetwyler et al. 2007; Gaynor et al. 2021; Woolliams et al. 2015), where the relationship of candidates is determined from genomic data.

In genomic selection of hybrids based on predicted values of GCA of their parents, we suggest to select different proportions in the two parent populations, if these differ substantially in their population parameters such as the GCA variances and/or prediction accuracy of GCA effects.

Appendix 1: Response to truncation selection in two sets and composition of the selected set

In our notation, we use $\varphi(x)$ and $\Phi(x)$ to denote the probability density function and cumulative distribution function of the standard normal distribution $N(0, 1)$, respectively; $\alpha(x) = 1 - \Phi(x)$ and $i_{\alpha(x)}$ denote the selected proportion and selection intensity, respectively, when applying threshold x to a standard normal distribution $N(0, 1)$.

Our assumptions for the mathematical derivations are:

1. Z is a Bernoulli distributed variable that indicates the origin of a candidate C from set Π_1 or Π_2, where $Z = 1$ with probability $\pi_1 = \frac{|\Pi_1|}{|\Pi_1 \cup \Pi_2|}$ for $C \in \Pi_1$ and $Z = 2$ with probability $\pi_2 = 1 - \pi_1$ for $C \in \Pi_2$.
2. X is the random variable for the SC with a conditional distribution $X|_{Z=1} \sim N(\mu_1, \sigma_1^2)$ for $C \in \Pi_1$ and $X|_{Z=2} \sim N(\mu_2, \sigma_2^2)$ for candidates $C \in \Pi_2$.
3. Candidates C from set Π_1 and Π_2 are selected based on their SC surpassing the respective thresholds t_1 and

t_2, yielding the sets Γ_1 and Γ_2 of selected candidates. Choosing thresholds t_1 and t_2 is equivalent to selecting proportions α_1 and α_2 of top candidates from Π_1 and Π_2, respectively, where

$$\begin{aligned}\alpha_1 &= P\left[X > t_1 | C \in \Pi_1\right] = P\left[X|_{Z=1} > t_1\right] \\ &= \left[1 - \Phi\left(\frac{t_1 - \mu_1}{\sigma_1}\right)\right] = \alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right) \\ \alpha_2 &= P\left[X > t_2 | C \in \Pi_2\right] = P\left[X|_{Z=2} > t_2\right] \\ &= \left[1 - \Phi\left(\frac{t_2 - \mu_2}{\sigma_2}\right)\right] = \alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right).\end{aligned} \tag{16}$$

Our subsequent derivations are based on thresholds as dealing with proportions would further complicate the already complex algebra. Applying the theorem of total probability and defining $\boldsymbol{\xi} = \left(\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2\right)$, we obtain for the total proportion of candidates selected from$\Pi_1 \cup \Pi_2$:

$$\begin{aligned}&\alpha_{\mathrm{Tot}}\left(t_1, t_2, \boldsymbol{\xi}\right) \\ &= P\left[X > t_1 | C \in \Pi_1\right] \cdot P\left[C \in \Pi_1\right] \\ &\quad + P\left[X > t_2 | C \in \Pi_2\right] \cdot P\left[C \in \Pi_2\right] \\ &= P\left[X|_{Z=1} > t_1\right] P[Z = 1] + P\left[X|_{Z=2} > t_2\right] P[Z = 2] \\ &= \alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right)\pi_1 + \alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right)\pi_2\end{aligned} \tag{17}$$

Using Bayes' formula, the proportion of candidates from Π_1 in the entire set $\Gamma_1 \cup \Gamma_2$ of selected candidates is

$$\begin{aligned}\gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right) &= \frac{P\left[X|_{Z=1} > t_1\right] P[Z = 1]}{P\left[\left[X|_{Z=1} > t_1\right] \vee \left[X|_{Z=2} > t_2\right]\right]} \\ &= \frac{\alpha\left(\frac{t_1 - \mu_1}{\sigma_{\tilde{u}_1}}\right)\pi_1}{\alpha_{\mathrm{Tot}}\left(t_1, t_2, \boldsymbol{\xi}\right)} = \frac{|\Gamma_1|}{|\Gamma_1 \cup \Gamma_2|} \text{ and} \\ \gamma_2\left(t_1, t_2, \boldsymbol{\xi}\right) &= 1 - \gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right).\end{aligned} \tag{18}$$

The expectation of the SC for the candidates in Γ_1 and Γ_2 selected from Π_1 and Π_2, respectively, is

$$\begin{aligned}E\left[X|_{Z=1} > t_1\right] &= \sigma_1 i_{\alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right)} + \mu_1 \text{ and} \\ E\left[X|_{Z=2} > t_2\right] &= \sigma_2 i_{\alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right)} + \mu_2,\end{aligned} \tag{19}$$

where $\sigma i_{\alpha\left(\frac{t-\mu}{\sigma}\right)}$ is the selection differential under truncation selection with threshold t in a normal distribution $N\left(\mu, \sigma^2\right)$, and $i_{\alpha(x)}$ is the selection intensity (Falconer and Mackay 1996, p. 189).

Thus, we obtain the expectation of the selected candidates in $\Gamma_1 \cup \Gamma_2$ as

$$\begin{aligned}&E\left[X|\left[Z = 1 \wedge X|_{Z=1} > t_1\right] \vee \left[Z = 2 \wedge X|_{Z=2} > t_2\right]\right] \\ &= E\left[X|_{Z=1} > t_1\right] P\left[X|_{Z=1} > t_1 | \left[X|_{Z=1} > t_1\right] \vee \left[X|_{Z=2} > t_2\right]\right] \\ &\quad + E\left[X|_{Z=2} > t_2\right] P\left[X|_{Z=2} > t_2 | \left[X|_{Z=1} > t_1\right] \vee \left[X|_{Z=2} > t_2\right]\right] \\ &= \left[\sigma_1 i_{\alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right)} + \mu_1\right]\gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right) \\ &\quad + \left[\sigma_2 i_{\alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right)} + \mu_2\right]\gamma_2\left(t_1, t_2, \boldsymbol{\xi}\right)\end{aligned} \tag{20}$$

Since the mean of unselected candidates in $\Pi_1 \cup \Pi_2$ is obtained as

$$\begin{aligned}&E[X|[Z = 1] \vee [Z = 1]] = E[X|Z = 1]P[Z = 1] \\ &+ E[X|Z = 2]P[Z = 2] = \mu_1\pi_1 + \mu_2\pi_2\end{aligned}$$

we get for the change in the mean of the SC as a result of truncation selection (= selection differential)

$$\begin{aligned}\mathrm{SD}_{\mathrm{Tot}}\left(t_1, t_2, \boldsymbol{\xi}\right) = \sigma_1 i_{\alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right)}\gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right) + \sigma_2 i_{\alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right)}\gamma_2\left(t_1, t_2, \boldsymbol{\xi}\right) \\ + \mu_1\left[\gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right) - \pi_1\right] + \mu_2\left[\gamma_2\left(t_1, t_2, \boldsymbol{\xi}\right) - \pi_2\right]\end{aligned} \tag{21}$$

Assuming the regression coefficient of the TGV on the SC is b_1 in Π_1 and b_2 in Π_2 and applying the breeders' equation for each set, we get for the total selection response in $\Pi_1 \cup \Pi_2$ under truncation selection with thresholds t_1 and t_2:

$$\begin{aligned}&\Delta G_{\mathrm{Tot}}\left(t_1, t_2, \boldsymbol{\xi}, b_1, b_2\right) \\ &= \Delta G_1\left(\left(t_1 - \mu_1\right), \sigma_1, b_1\right)\gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right) \\ &\quad + \Delta G_2\left(\left(t_2 - \mu_2\right), \sigma_2, b_2\right)\gamma_2\left(t_1, t_2, \boldsymbol{\xi}\right) \\ &\quad + \mu_1\left[\gamma_1\left(t_1, t_2, \boldsymbol{\xi}\right) - \pi_1\right] \\ &\quad + \mu_2\left[\gamma_2\left(t_1, t_2, \boldsymbol{\xi}\right) - \pi_2\right]\end{aligned} \tag{22}$$

where $\Delta G_1\left(\left(t_1 - \mu_1\right), \sigma_1, b_1\right) = b_1\sigma_1 i_{\alpha\left(\frac{t_1 - \mu_1}{\sigma_1}\right)}$ and $\Delta G_2\left(\left(t_2 - \mu_2\right), \sigma_2, b_2\right) = b_2\sigma_2 i_{\alpha\left(\frac{t_2 - \mu_2}{\sigma_2}\right)}$ refer to the selection response realized in set Π_1 and Π_2, respectively.

Appendix 2: Maximizing selection response by optimal choice of selection thresholds

To determine the maximum of the selection response $G_{\mathrm{Tot}}\left(t_1, t_2, \boldsymbol{\xi}, b_1, b_2\right)$ as a function of the thresholds t_1 and t_2, we use a Lagrange multiplier approach. We start with some basic properties of the normal distribution $N(0, 1)$ with pdf $\varphi(x)$ and cdf $\Phi(x)$. Let $i_{\alpha(x)} = \frac{\varphi(x)}{\alpha(x)}$ and $\alpha(x) = \int_x^{\infty}\varphi(z)dz = 1 - \Phi(x)$ be the selection intensity and selected proportion, respectively, then we have

$\frac{\partial\varphi(x)}{\partial x} = -x\varphi(x)$ and $\frac{\partial\alpha(x)}{\partial x} = -\varphi(x)$. Hence, setting $x = \frac{t-\mu}{\sigma}$, we obtain.

$$\frac{\partial\varphi\left(\frac{t-\mu}{\sigma}\right)}{\partial t} = -\left(\frac{t-\mu}{\sigma}\right)\varphi\left(\frac{t-\mu}{\sigma}\right)\frac{1}{\sigma}. \tag{23}$$

$$\frac{\partial\alpha\left(\frac{t-\mu}{\sigma}\right)}{\partial t} = -\varphi\left(\frac{t-\mu}{\sigma}\right)\frac{1}{\sigma}. \tag{24}$$

Defining for given values of $\boldsymbol{\xi}, b_1, b_2$ and $\alpha_T \in (0,1)$:

$$\beta_1(t_1) = \alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)\pi_1,\ \beta_2(t_2) = \alpha\left(\frac{t_2-\mu_2}{\sigma_2}\right)\pi_2,$$
$$g(t_1,t_2) = \beta_1(t_1) + \beta_2(t_2) - \alpha_T$$

and

$$\begin{aligned} f(t_1,t_2) &= \Delta G_1\big((t_1-\mu_1),\sigma_1,b_1\big)\beta_1(t_1) \\ &\quad + \Delta G_2\big((t_2-\mu_2),\sigma_2,b_2\big)\beta_2(t_2) \\ &\quad + \mu_1\big(\beta_1(t_1)-\pi_1\alpha_T\big) + \mu_2\big(\beta_2(t_2)-\pi_2\alpha_T\big), \end{aligned}$$

we get the Lagrangian function

$$L(t_1,t_2,\lambda) = f(t_1,t_2) + \lambda g(t_1,t_2) \tag{25}$$

Thus, for given values of $\boldsymbol{\xi}, b_1, b_2$ and α, we obtain a necessary condition for the maximum of $\Delta G_{\mathrm{Tot}}(t_1,t_2,\boldsymbol{\xi},b_1,b_2) = \frac{1}{\alpha_T} f(t_1,t_2)$ under the side condition $g(t_1,t_2)=0$ by analyzing the gradient $\nabla L(t_1,t_2,\lambda)$.

We have

$$\begin{aligned} \Delta G_1\big((t_1-\mu_1),\sigma_1,b_1\big)\beta_1(t_1) &= b_1\sigma_1\frac{\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)}{\alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)}\alpha\left(\frac{t_1-\mu_1}{\sigma_1}\right)\pi_1 \\ &= b_1\sigma_1\pi_1\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right) \end{aligned}$$

and

$$\Delta G_2\big((t_2-\mu_2),\sigma_2,b_2\big)\beta_2(t_2) = b_2\sigma_2\pi_2\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right).$$

Thus,

$$\begin{aligned} &\frac{\partial(\Delta G_1\big((t_1-\mu_1),\sigma_1,b_1\big)\beta_1(t_1)}{\partial t_1} \\ &= -b_1\sigma_1\pi_1\left(\frac{t_1-\mu_1}{\sigma_1}\right)\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1} \\ &= -b_1\pi_1(t_1-\mu_1)\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1} \end{aligned}$$

$$\begin{aligned} &\frac{\partial\Delta G_2\big((t_2-\mu_2),\sigma_2,b_2\big)\beta_2(t_2)}{\partial t_1} = 0, \\ &\frac{\partial\big(\mu_1\big(\beta_1(t_1)-\pi_1\alpha_{\mathrm{Tot}}\big)\big)}{\partial t_1} = -\mu_1\pi_1\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1}, \\ &\frac{\partial\big(\mu_2\big(\beta_2(t_2)-\pi_2\alpha_{\mathrm{Tot}}\big)\big)}{\partial t_1} = 0, \end{aligned}$$

$$\frac{\partial\lambda g(t_1,t_2)}{\partial t_1} = -\lambda\pi_1\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1}.$$

Likewise,

$$\begin{aligned} &\frac{\partial(\Delta G_1\big((t_1-\mu_1),\sigma_1,b_1\big)\beta_1(t_1)}{\partial t_2} = 0, \\ &\frac{\partial\Delta G_2\big((t_2-\mu_2),\sigma_2,b_2\big)\beta_2(t_2)}{\partial t_2} = -b_2\pi_2(t_2-\mu_2)\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right)\frac{1}{\sigma_2}, \end{aligned}$$

$$\begin{aligned} &\frac{\partial\big(\mu_1\big(\beta_1(t_1)-\pi_1\alpha_T\big)\big)}{\partial t_2} = 0, \\ &\frac{\partial\big(\mu_2\big(\beta_2(t_2)-\pi_2\alpha_T\big)\big)}{\partial t_2} = -\mu_2\pi_2\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right)\frac{1}{\sigma_2}, \\ &\frac{\partial\lambda g(t_1,t_2)}{\partial t_2} = -\lambda\pi_2\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right)\frac{1}{\sigma_2}. \end{aligned}$$

$$\begin{aligned} &\frac{\partial(\Delta G_1\big((t_1-\mu_1),\sigma_1,b_1\big)\beta_1(t_1)}{\partial\lambda} = 0, \\ &\frac{\partial\Delta G_2\big((t_2-\mu_2),\sigma_2,b_2\big)\beta_2(t_2)}{\partial\lambda} = 0, \\ &\frac{\partial(\mu_1\big(\beta_1(t_1)-\alpha_T\big))}{\partial\lambda} = 0, \\ &\frac{\partial(\mu_2\big(\beta_1(t_1)-\pi_1\alpha_T\big))}{\partial\lambda} = 0 \end{aligned}$$

and

$$\frac{\partial\lambda g(t_1,t_2)}{\partial\lambda} = g(t_1,t_2)$$

Thus, we get

$$\begin{aligned} \frac{\partial L(t_1,t_2,\lambda)}{\partial t_1} &= -b_1\pi_1(t_1-\mu_1)\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1} \\ &\quad - \mu_1\pi_1\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1} - \lambda\pi_1\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\frac{1}{\sigma_1} \\ &= -\frac{\pi_1}{\sigma_1}\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right)[(b_1(t_1-\mu_1)+\mu_1+\lambda] \end{aligned} \tag{26}$$

$$\frac{\partial L(t_1,t_2,\lambda)}{\partial t_2} = -\frac{\pi_2}{\sigma_2}\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right)[(b_2(t_2-\mu_2)+\mu_2+\lambda] \tag{27}$$

$$\frac{\partial L(t_1,t_2,\lambda)}{\partial \lambda} = g(t_1,t_2) \tag{28}$$

Setting $\nabla L(t_1,t_2,\lambda) = (0,0,0)$ and using that $\frac{\pi_1}{\sigma_1}\varphi\left(\frac{t_1-\mu_1}{\sigma_1}\right) > 0$ and $\frac{\pi_2}{\sigma_2}\varphi\left(\frac{t_2-\mu_2}{\sigma_2}\right) > 0$, we obtain from Eqs. 26 and 27 the necessary conditions: -$\lambda = b_1(t_1-\mu_1)+\mu_1$ and - $\lambda = b_2(t_2-\mu_2)+\mu_2$ or equivalently $b_1(t_1-\mu_1)+\mu_1 = b_2(t_2-\mu_2)+\mu_2$. Thus, the solutions t_1^*, t_2^* must fulfill the following conditions:

$$\begin{aligned} t_1 &= \frac{b_2(t_2-\mu_2)+\mu_2-\mu_1}{b_1}+\mu_1 \text{ or equivalently} \\ t_2 &= \frac{b_1(t_1-\mu_1)+\mu_1-\mu_2}{b_2}+\mu_2 \end{aligned} \tag{29}$$

and

$$\Phi\left(\frac{t_1-\mu_1}{\sigma_1}\right)\pi_1 + \Phi\left(\frac{t_2-\mu_2}{\sigma_2}\right)\pi_2 = 1-\alpha_T \tag{30}$$

For the special case of BLUPs, where $b_1 = b_2 = 1.0$, we have $t_1^* = t_2^*$. (31)

For BLUEs and $\mu_1 = \mu_2$, we get $t_1^* - \mu_1 = \frac{b_2(t_2^*-\mu_1)}{b_1}$. (32)

Defining $t_1^{**} = t_1^* - \mu_1$ and $t_2^{**} = t_2^* - \mu_1$, Eqs. 29 and 30 are equivalent to $t_1^{**} = \frac{b_2 t_2^{**}}{b_1}$ and $\Phi\left(\frac{t_1^{**}}{\sigma_1}\right)\pi_1 + \Phi\left(\frac{t_2^{**}}{\sigma_2}\right)\pi_2 = 1-\alpha_T$, which would be obtained if we assume $\mu_1 = 0$. The improvement in $\Delta G_{Tot}(t_1^*, t_2^*, \xi, b_1, b_2)$ relative to $\Delta G_{Tot}(t_1^i, t_2^i, \xi, b_1, b_2)$, where $t_1^i = t_2^i$ such that $\alpha_{Tot}(t_1^i, t_2^i, \xi) = \alpha_T$ or equivalently $\Phi\left(\frac{t_1^i-\mu_2}{\sigma_1}\right)\pi_1 + \Phi\left(\frac{t_2^i-\mu_2}{\sigma_2}\right)\pi_2 = 1-\alpha_T$, can be expressed as the ratio $\Psi_{Tot}(\alpha_T, \xi, b_1, b_2)$ defined in Eq. 10.

Appendix 3: Maximizing the total selection response for factorials in hybrid breeding

Since the SC is expected to provide unbiased estimates for the GCA effects in each parent population, we have $\mu_1 = \mu_2 = 0$. Additionally, we assume that b_1 and b_2 are the regression coefficients of the regression of GCA effects on the SC in Π_1 and Π_2, respectively, and define $\boldsymbol{\theta} = (\sigma_1, \sigma_2, b_1, b_2)$. We select proportions $\alpha_1 = \frac{|\Gamma_1|}{|\Pi_1|}$ and $\alpha_2 = \frac{|\Gamma_2|}{|\Pi_2|}$ from Π_1 and Π_2, respectively, by applying corresponding thresholds $t_1 = \sigma_1\Phi^{-1}(1-\alpha_1)$ and $t_2 = \sigma_2\Phi^{-1}(1-\alpha_2)$. If the selected lines are mated in the form of a complete factorial design to produce the hybrids tested in the next step of the breeding program, the proportion of selected hybrids from the total set of possible hybrids is

$$\alpha_{Hyb}(t_1,t_2,\boldsymbol{\theta}) = \alpha\left(\frac{t_1}{\sigma_1}\right)\times\alpha\left(\frac{t_2}{\sigma_2}\right). \tag{33}$$

The expected selection response in the hybrids among lines from Π_1 and Π_2, which were selected for their GCA in cross-combinations with genotypes from the other population, can be obtained using Eq. 6 as follows

$$\Delta G_{Hyb}(t_1,t_2,\boldsymbol{\theta}) = \Delta G_1(t_1,\sigma_1,b_1) + \Delta G_2(t_2,\sigma_2,b_2), \tag{34}$$

with

$$\Delta G_1(t_1,\sigma_1,b_1) = b_1\sigma_1\frac{\varphi\left(\frac{t_1}{\sigma_1}\right)}{\alpha\left(\frac{t_1}{\sigma_1}\right)} = \frac{1}{\alpha_{Hyb}(t_1,t_2,\boldsymbol{\theta})}b_1\sigma_1\varphi\left(\frac{t_1}{\sigma_1}\right)\alpha\left(\frac{t_2}{\sigma_2}\right) \tag{35}$$

and

$$\Delta G_2(t_2,\sigma_2,b_2) = \frac{1}{\alpha_{Hyb}(t_1,t_2,\boldsymbol{\theta})}b_2\sigma_2\varphi\left(\frac{t_2}{\sigma_2}\right)\alpha\left(\frac{t_1}{\sigma_1}\right) \tag{36}$$

Defining for given values of $\boldsymbol{\theta}$ and $\alpha_H \in (0,1)$:

$$\beta_1(t_1) = \alpha\left(\frac{t_1}{\sigma_1}\right),\ \beta_2(t_2) = \alpha\left(\frac{t_2}{\sigma_2}\right),\ g(t_1,t_2) = \beta_1(t_1)\beta_2(t_2) - \alpha_H$$

and

$$f(t_1,t_2) = b_1\sigma_1\varphi\left(\frac{t_1}{\sigma_1}\right)\alpha\left(\frac{t_2}{\sigma_2}\right) + b_2\sigma_2\varphi\left(\frac{t_2}{\sigma_2}\right)\alpha\left(\frac{t_1}{\sigma_1}\right),$$

We get the Lagrangian function

$$L(t_1,t_2,\lambda) = f(t_1,t_2) + \lambda g(t_1,t_2) \tag{37}$$

Thus, for given values of $\boldsymbol{\theta}$ and α_H, we obtain a necessary condition for the maximum of $\Delta G_{Hyb}(t_1,t_2,\boldsymbol{\theta}) = \frac{1}{\alpha_H}f(t_1,t_2)$ under the side condition $g(t_1,t_2)=0$ by analyzing the gradient $\nabla L(t_1,t_2,\lambda)$. We have

$$\begin{aligned} \frac{\partial L(t_1,t_2\lambda)}{\partial t_1} &= -b_1\sigma_1\frac{t_1}{\sigma_1}\varphi\left(\frac{t_1}{\sigma_1}\right)\frac{1}{\sigma_1}\alpha\left(\frac{t_2}{\sigma_2}\right) \\ &\quad + b_2\sigma_2\varphi\left(\frac{t_2}{\sigma_2}\right)\times\left(-\varphi\left(\frac{t_1}{\sigma_1}\right)\frac{1}{\sigma_1}\right) \\ &\quad + \lambda\left(-\varphi\left(\frac{t_1}{\sigma_1}\right)\frac{1}{\sigma_1}\alpha\left(\frac{t_2}{\sigma_2}\right)\right) \\ &= \frac{-\varphi\left(\frac{t_1}{\sigma_1}\right)\alpha\left(\frac{t_2}{\sigma_2}\right)}{\sigma_1}\left[b_1t_1 + b_2\sigma_2 i_{\alpha\left(\frac{t_2}{\sigma_2}\right)} + \lambda\right] \end{aligned} \tag{38}$$

$$\frac{\partial L(t_1,t_2\lambda)}{\partial t_2} = \frac{-\varphi\left(\frac{t_2}{\sigma_2}\right)\alpha\left(\frac{t_1}{\sigma_1}\right)}{\sigma_2}\left[b_1\sigma_1 i_{\alpha\left(\frac{t_1}{\sigma_1}\right)} + b_2t_2 + \lambda\right] \tag{39}$$

$$\frac{\partial L(t_1,t_2\lambda)}{\partial \lambda} = \alpha\left(\frac{t_1}{\sigma_1}\right)\alpha\left(\frac{t_2}{\sigma_2}\right) - \alpha_{\mathrm{H}} \tag{40}$$

Thus, from $\left(\frac{\nabla L(t_1,t_2,\lambda)}{\partial t_1}, \frac{\nabla L(t_1,t_2,\lambda)}{\partial t_2}, \frac{\nabla L(t_1,t_2,\lambda)}{\partial \lambda}\right) = (0,0,0)$, we get the necessary conditions $-\lambda = b_1 t_1 + b_2\sigma_2 i_{\alpha\left(\frac{t_2}{\sigma_2}\right)}$ and $-\lambda = b_1\sigma_1 i_{\alpha\left(\frac{t_1}{\sigma_1}\right)} + b_2 t_2$ or equivalently

$$\begin{aligned} b_1 t_1 - b_2 t_2 &= b_1\sigma_1 i_{\alpha\left(\frac{t_1}{\sigma_1}\right)} - b_2\sigma_2 i_{\alpha\left(\frac{t_2}{\sigma_2}\right)} \\ &= \Delta G_1(t_1,\sigma_1,b_1) - \Delta G_1(t_1,\sigma_1,b_1) \end{aligned} \tag{41}$$

and

$$\alpha\left(\frac{t_1}{\sigma_1}\right)\alpha\left(\frac{t_2}{\sigma_2}\right) = \alpha_{\mathrm{H}} \tag{42}$$

Numerical solutions $\left(t_1^o, t_2^o\right)$ for Eqs. 41 and 42 can be obtained using mathematical software such as Mathematica, from which we get $\alpha_1^o = \alpha\left(\frac{t_1^o}{\sigma_1}\right)$ and $\alpha_2^o = \alpha\left(\frac{t_2^o}{\sigma_1}\right)$ and $\Delta G_{\mathrm{Hyb}}\left(t_1^o, t_2^o, \boldsymbol{\theta}\right)$. This maximum can be compared with the selection response $\Delta G_{\mathrm{Hyb}}\left(t_1^e, t_2^e, \boldsymbol{\theta}\right)$ for $t_1^e = \sigma_1\Phi^{-1}\left(1-\sqrt{\alpha_{\mathrm{H}}}\right)$ and $t_2^e = \sigma_2\Phi^{-1}\left(1-\sqrt{\alpha_{\mathrm{H}}}\right)$, i.e., when an equal proportion $\alpha^e = \sqrt{\alpha_{\mathrm{H}}}$ of lines is selected for GCA in each parent population. For selection based on BLUPs, where $b_1 = b_2 = 1$, Eq. 40 simplifies to

$$t_1 - t_2 = \sigma_1 i_{\alpha\left(\frac{t_1}{\sigma_1}\right)} - \sigma_2 i_{\alpha\left(\frac{t_2}{\sigma_2}\right)} \tag{43}$$

which corresponds to the difference in the selection differentials in Π_1 and Π_2.

Appendix 4: Regression equation of true genetic values (TGVs) on their BLUPs

We consider the ordinary mixed linear model studied by Henderson (1975)

$$y = X\beta + Zu + e,$$

where y is an $n \times 1$ observation vector, X a known $n \times p$ matrix with full column rank p, β an unknown fixed vector, and Z is a known $n \times q$ matrix. u and e are unobservable random vectors with null means and

$$\mathrm{var}\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix},$$

where G and R are both nonsingular and known matrices.

Let $\hat{u}$ be the BLUP of u calculated as described by Henderson (1975) and denote $\Sigma_{22} = \mathrm{var}(\hat{u})$ and $\Sigma_{12} = \mathrm{cov}(u, \hat{u})$. Then, according to Henderson (1975), $\Sigma_{22} = \Sigma_{12}$ so that we have $\Sigma_{12}\Sigma_{22}^{-1} = I$, if $\boldsymbol{\Sigma}_{22}$ can be inverted,

In the case, where $\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix}$ follows a multivariate normal distribution so that $\hat{\boldsymbol{u}}$ also follows a normal distribution, the regression function of $\boldsymbol{u}$ on $\hat{\boldsymbol{u}}$ is linear (Anderson 1958, p. 29) and can be expressed as $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\hat{\boldsymbol{u}} = \boldsymbol{I}\hat{\boldsymbol{u}}$.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00122-024-04592-2.

Acknowledgements The authors are indebted to Prof. Daniel Gianola for critical reading and helpful suggestions on an earlier version of the manuscript. ChatGPT 3.5 by OpenAI has been used to improve the grammar and style of this paper.

Authors contribution statement AEM conceived the study and developed the theory. AEM and AJM developed jointly the Mathematica programs and the figures. AEM wrote the manuscript with support from RF and CCS. All authors discussed and interpreted the results, read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by intra-mural funds of the Technical University of Munich. Open access was enabled and organized by Projekt Deal.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. AEM is editor-in-chief and CCS is member of the editorial board of Theor. Appl. Genetics.

Ethical standard The authors declare that their work complies with the current laws of Germany.

References

Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York

Auinger H-J, Lehermeier C, Gianola D, Mayer M, Melchinger AE, da Silva S, Knaak C, Ouzunova M, Schön C-C (2021) Calibration and validation of predicted genomic breeding values in an advanced cycle maize population. Theor Appl Genet 134:3069–3081

Barbosa PAM, Fritsche-Neto R, Andrade MC, Petroli CD, Burgueño J, Galli G, Willcox MC, Sonder K, Vidal-Martínez VA, Sifuentes-Ibarra E (2021) Introgression of maize diversity for drought

tolerance: subtropical maize landraces as source of new positive variants. Front Plant Sci 12:691211
Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20–25
Bernardo R (1996) Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:50–56
Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press, Woodbury
Böhm J, Schipprack W, Utz HF, Melchinger AE (2017) Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. Theor Appl Genet 130:861–873
Bonnett D, Li Y, Crossa J, Dreisigacker S, Basnet B, Pérez-Rodríguez P, Alvarado G, Jannink J-L, Poland J, Sorrells M (2022) Response to early generation genomic selection for yield in wheat. Front Plant Sci 12:718611
Brauner PC, Müller D, Molenaar WS, Melchinger AE (2019) Genomic prediction with multiple biparental families. Theor Appl Genet 133:133–147
Brotherstone S, Hill W (1986) Heterogeneity of variance amongst herds for milk production. Anim Sci 42:297–303
Bulmer MG (1980) The mathematical theory of quantitative genetics. Clarendon Press, New York
Chaikam V, Molenaar W, Melchinger AE, Boddupalli PM (2019) Doubled haploid technology for line development in maize: technical advances and prospects. Theor Appl Genet 132:3227–3243
Clark SA, Hickey JM, Daetwyler HD, van der Werf JH (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44:1–9
Cochran W (1951) Improvement by means of selection. In: Proceedings of the second Berkeley symposium on mathematical statistics and probability, pp 449–470
Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection. J Anim Breed Genet 124:369–376
Falconer D, Mackay T (1996) Introduction to quantitative genetics. Longman Group, Essex
Fernando R, Gianola D (1986) Optimal properties of the conditional mean as a selection criterion. Theor Appl Genet 72:822–825
Garrick D, Van Vleck LD (1987) Aspects of selection for performance in several environments with heterogeneous variances. J Anim Sci 65:409–421
Gaynor RC, Gorjanc G, Hickey JM (2021) AlphaSimR: an R package for breeding program simulations. G3 11:jkaa017
Goffinet B (1983) Selection on selected records. Génét Sélect Évol 15:91–98
Habier D, Fernando RL, Dekkers J (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397
Hartl DL, Clark AG, Clark AG (1997) Principles of population genetics. Sinauer Associates, Sunderland
Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447
Henderson C (1990) Statistical methods in animal improvement: historical overview. In: Advances in statistical methods for genetic improvement of livestock. Springer, pp 2–14
Hill W (1984) On selection among groups with heterogeneous variance. Anim Sci 39:473–477
Hölker AC, Mayer M, Presterl T, Bolduan T, Bauer E, Ordas B, Brauner PC, Ouzunova M, Melchinger AE, Schön C-C (2019) European maize landraces made accessible for plant breeding and genome-based studies. Theor Appl Genet 132:3333–3345
Kennedy B, Sorenson D (1988) Properties of mixed model methods for prediction of genetic merit under different genetic models in selected and nonselected populations. In: Second international conference on quantitative genetics, Raleigh. Sinauer Associates, pp 47–56
Lehermeier C, Krämer N, Bauer E, Bauland C, Camisan C, Campo L, Flament P, Melchinger AE, Menz M, Meyer N (2014) Usefulness of multiparental populations of maize (*Zea mays* L.) for genome-based prediction. Genetics 198:3–16
Lian L, Jacobson A, Zhong S, Bernardo R (2014) Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci 54:1514–1522
Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland
Mayer M, Unterseer S, Bauer E, de Leon N, Ordas B, Schön CC (2017) Is there an optimum level of diversity in utilization of genetic resources? Theor Appl Genet 130:2283–2295
Melchinger AE, Fernando R, Stricker C, Schön CC, Auinger HJ (2023) Genomic prediction in hybrid breeding: I. Optimizing the training set design. Theor Appl Genet 136:176
Melchinger AE, Frisch M (2023) Genomic prediction in hybrid breeding: II. Reciprocal recurrent genomic selection with full-sib and half-sib families. Theor Appl Genet 136:203
Melchinger AE, Posselt UK (2013) Biotechnologie und Züchtung. In: Lütke-Entrup NS, Schwarz FJ, Heilmann H (eds) Handbuch Mais. DLG Verlag, Frankfurt, M, pp 53–64
Piepho H, Möhring J, Melchinger A, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228
Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z (2017) Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant 10:1047–1064
Riedelsheimer C, Melchinger AE (2013) Optimizing the allocation of resources for genomic selection in one breeding cycle. Theor Appl Genet 126:2835–2848
Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink J-L, Melchinger AE (2013) Genomic predictability of interconnected Bi-parental maize populations. Genetics 194:493–503
Robert P, Auzanneau J, Goudemand E, Oury F-X, Rolland B, Heumez E, Bouchet S, Le Gouis J, Rincent R (2022) Phenomic selection in wheat breeding: identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. Theor Appl Genet 135:895–914
Schnell F (1982) A synoptic study of the methods and categories of plant breeding
Schrag T, Melchinger A, Sørensen A, Frisch M (2006) Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor Appl Genet 113:1037–1047
Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE (2018) Beyond genomic prediction: combining different types of omics data can improve prediction of hybrid performance in maize. Genetics 208:1373–1385
Seifert F, Thiemann A, Schrag TA, Rybka D, Melchinger AE, Frisch M, Scholten S (2018) Small RNA-based prediction of hybrid performance in maize. BMC Genom 19:1–14
Seye A, Bauland C, Charcosset A, Moreau L (2020) Revisiting hybrid breeding designs using genomic predictions: simulations highlight the superiority of incomplete factorials between segregating families over topcross designs. Theor Appl Genet 133:1995–2010
Sorenson D, Gianola D (2004) Likelihood, bayesian, and MCMC methods in quantitative genetics. Springer, New York
Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343–1355

Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J, Rey M-D, Asyraf Md, Hatta M, Hinchliffe A, Steed A, Reynolds D (2018) Speed breeding is a powerful tool to accelerate crop research and breeding. Nat Plants 4:23–29

Weiß TM, Zhu X, Leiser WL, Li D, Liu W, Schipprack W, Melchinger AE, Hahn V, Würschum T (2022) Unraveling the potential of phenomic selection within and among diverse breeding material of maize (Zea mays L.). G3 12:jkab445

Westhues M, Schrag TA, Heuer C, Thaller G, Utz HF, Schipprack W, Thiemann A, Seifert F, Ehret A, Schlereth A (2017) Omics-based hybrid prediction in maize. Theor Appl Genet 130:1927–1939

Westhues M, Heuer C, Thaller G, Fernando R, Melchinger AE (2019) Efficient genetic value prediction using incomplete omics data. Theor Appl Genet 132:1211–1222

Wilde P, Menzel J, Schmiedchen B (2003) Estimation of general and specific combining ability variances and their implications on hybrid rye breeding. Plant Breed Seed Sci 47:89–98

Wilde P, Miedaner T (2021) Hybrid rye breeding. In: The rye genome, pp 13–41

Wolfram S (1999) The MATHEMATICA® book, version 4. Cambridge University Press, Cambridge

Woolliams J, Berg P, Dagnachew B, Meuwissen T (2015) Genetic contributions and their optimization. J Anim Breed Genet 132:89–99

Zenke-Philippi C, Frisch M, Thiemann A, Seifert F, Schrag T, Melchinger AE, Scholten S, Herzog E (2017) Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme. Plant Breed 136:331–337