Received: 3 July 2024 | Accepted: 11 September 2024 | Published online: 9 October 2024

DOI: 10.1002/csc2.21384

Crop Science

ORIGINAL ARTICLE

Crop Breeding & Genetics

Elite germplasm introduction, training set composition, and genetic optimization algorithms effect on genomic selection-based breeding programs

Roberto Fritsche-Neto[1] | **Rafael Massahiro Yassue**[2] | **Allison Vieira da Silva**[3] | **Melina Prado**[3] | **Júlio César DoVale**[4]

[1]H. Rouse Caffey Rice Research Station, LSU AgCenter, Rayne, Louisiana, USA

[2]GDM Seeds, Campinas, São Paulo, Brazil

[3]Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba, São Paulo, Brazil

[4]Department of Crop Science, Federal University of Ceará, Fortaleza, Ceará, Brazil

Correspondence
Roberto Fritsche-Neto, H. Rouse Caffey Rice Research Station, LSU AgCenter, Rayne, LA, USA.
Email: rfneto@agcenter.lsu.edu

Assigned to Associate Editor Alexander Lipka.

Funding information
Louisiana Rice Research Board; GDM Seeds, Grant/Award Number: GR-00013425

Abstract

In genomic selection (GS), the prediction accuracy is heavily influenced by the composition of the training set (TS). Currently, two primary strategies for building TS are used: one involves accumulating historical phenotypic records from multiple years, while the other is the "test-and-shelf" approach. Additionally, studies have suggested that optimizing TS composition using genetic algorithms can improve the accuracy of prediction models. Most breeders operate in open systems, introducing new genetic variability into their populations as needed. However, the impact of elite germplasm introduction in GS models remains unclear. Therefore, we conducted a case study in self-pollinated crops using stochastic simulations to understand the effects of elite germplasm introduction, TS composition, and its optimization in long-term breeding programs. Overall, introducing external elite germplasm reduces the prediction accuracy. In this context, test and shelf seem more stable regarding accuracy in dealing with introductions despite the origin and rate, being useful in programs where the introductions come from different sources over the years. Conversely, using historical data, if the introductions come from the same source over the cycles, this negative effect is reduced as long as the cycles and this approach become the best. Thus, it may support public breeding programs in establishing networks of collaborations where the exchange of germplasm will occur at a predefined rate and flow. In either case, the use of algorithms of optimization to trim the genetic variability does not bring a substantial advantage in the medium to long term.

Abbreviations: Fst, fixation Index; GEBV, genomic estimated breeding value; GPO, grandparent, parent, offspring; GS, genomic selection; LD, linkage disequilibrium; OTS, optimized TS; PEV, prediction error variance; PYT, preliminary yield trials; QTL, quantitative trait loci; QTN, quantitative trait nucleotide; SNP, single nucleotide polymorphism; TS, training set.

Plain Language Summary

In genomic selection (GS), prediction accuracy relies on how the training set (TS) is built. Two main methods are: 1. Historical data: Using data from multiple years. 2. Test and shelf: Regularly updating the TS with new data while keeping some old data. Our study found the following: (1) Elite germplasm: Adding new elite germplasm often reduces prediction accuracy due to increased genetic diversity. (2) TS strategies: Test and shelf: Provides stable accuracy, especially with varying germplasm sources. Historical data: It work well if new germplasm is from the same source over time. (3) Optimization algorithms: These offer limited long-term benefits. Implications: The test-and-shelf approach suits diverse sources, while historical data benefit consistent sources. Optimization algorithms are less crucial. For best results, focus on managing TS and planning germplasm introductions.

1 | INTRODUCTION

In breeding programs, genomic selection (GS) (Bernardo, 1994; Meuwissen et al., 2001) plays a pivotal role in predicting the genetic merit of individuals based on their DNA information. In this context, the accuracy of these predictions hinges significantly on the composition of the training set (TS) (Fernández-González et al., 2023; Sabadin et al., 2022). This set comprises genetic data from individuals with known phenotypic traits, such as yield, disease resistance, or quality. Based on this data, GS models estimate genetic markers' effects associated with desirable characteristics and use them to predict the performance of new individuals. Therefore, factors such as size, diversity of genetics and phenotypic traits, heritability and completeness of phenotypic data, and representativeness of the target population play crucial roles (Berro et al., 2019; Crossa et al., 2017; Isidro y Sánchez & Akdemir, 2021).

Overall, a diverse TS covering a wide array of genetic variations and phenotypic traits enables the algorithm to capture a broader spectrum of genetic effects and make more precise predictions across various populations and environments (Alemu et al., 2024). Conversely, a biased or incomplete TS may lead to unreliable predictions and hinder the performance of GS models. Therefore, ensuring the quality and diversity of the TS is vital for the success of predictions. This may involve meticulous selection of individuals for inclusion in the TS, ensuring representation across different populations, and consistently updating the dataset as new information becomes available (Sabadin et al., 2022).

Currently, the two main strategies for building TSs are based on accumulating historical phenotypic records from multiple years (Beyene et al., 2021; Rutkoski et al., 2015) and the test and shelf (Boyles et al., 2024). The advantage of the first approach is that the estimation of marker effects is based on multiple-year records, leading to a more accurate estimation of year effect and its interaction with genotypes. However, potential risks arise when considering the smaller connectivity between the training and testing sets and the need for joint analysis to adjust the phenotypic data (Bernal-Vasquez et al., 2017; Gonzalez et al., 2021). On the other hand, the test-and-shelf methodology consists of using part of the genotyped individuals from a larger population to be phenotyped, while the other part will be shelved and predicted only. The main advantage of this approach is that utilizing a portion of the same population from the TS to predict the other part ensures the highest level of connectivity between the training and testing sets. However, it is important to note that this method may overlook year–genotype interactions, mainly in real breeding programs, where every year is composed of a new batch of genotypes.

After defining the best strategy to build the TS, another aspect is the quality and repetitiveness of the data collected. In this context, optimizing the TS composition may enhance the accuracy of prediction models (Muleta et al., 2019). In other words, trim the genetic variability so only individuals who maximize the relationship with the target set will be considered in the TS (Akdemir & Isidro-Sánches, 2019). Overall, algorithms of optimization offer significant advantages such as increased prediction accuracy and cost reduction (Crossa et al., 2013; Heffner et al., 2009; Jarquín et al., 2014). Also, they address potential challenges related to computational complexity and overfitting.

Most studies showing the usefulness of genetic algorithms of optimization relied on empirical data (Fristche-Neto et al., 2018; de Freitas Mendonça & Fritsche-Neto, 2020; e Sousa et al., 2019). Consequently, they worked well to increase the prediction accuracy in the current dataset, with static haplotypes in terms of size and frequency. On the other hand, other authors performed studies using stochastic simulations (DoVale et al., 2022), showing that the use of these algorithms may significantly reduce the accuracy over breeding

cycles, where there will be changes in allele frequencies over time, and recombination, that will break the "big" haplotypes into small ones, and some of them may segregate without any maker tagging them, reducing the prediction accuracy over cycles.

In breeding, the primary goal is to effectively utilize genetic diversity to achieve improvement in specific traits while preserving genetic diversity for the coming cycles of selection (Meuwissen et al., 2020). For that, breeders aim to increase the frequency of targeted quantitative trait loci (QTL), despite the potential loss of genetic diversity at those and other sites. Moreover, the loss of genetic diversity through artificial selection is boosted by drift because breeding populations are often derived from a small number of parents (Juma et al., 2021; Y. Li, Shi et al., 2022). In this context, the decrease in diversity is only sustainable if the breeder can achieve significant long-term genetic gains (Allier et al., 2020; Meuwissen et al., 2020), defining clear strategies to preserve or "feed" the population with new genetic variability.

The identification and incorporation of valuable genetic resources into a breeding program is an important procedure that breeders have to maintain or recover genetic diversity in populations (Allier et al., 2020; Swarup et al., 2021). For that, the elite germplasm exchange between breeding programs or introduction is a useful approach that can be implemented to mitigate the problem of genetic diversity loss due to artificial selection or drift. For instance, the Island model for GS has shown that germplasm exchange from subpopulations derived from a large population might be a powerful tool to maintain genetic diversity while improving genetic gain (Yabe et al., 2016). Furthermore, most of the industry programs exchange germplasm intentionally. However, the impact of the one-way elite germplasm introduction in genomic prediction models has yet to be made clear. In this context, several aspects should be considered, such as haplotype phasing, overall genetic diversity, trait heritability, differences in marker and gene effects, and the extent and distribution between QTLs and markers (Kaler et al., 2022; Werner et al., 2020).

Given the abovementioned aspects, we compare several scenarios to understand the elite germplasm introduction, TS composition (test and shelf or historical data), and its optimization (using genetic optimization algorithms to trim the data) effect in long-term breeding programs. For that, we considered a study case in self-pollinated crops via stochastic simulations and evaluated effects in genetic gain, best line performance, genetic variance, genetic divergence between germplasm sources, and prediction accuracy. The choice for stochastic simulations is because comparing breeding strategies based only on field trials could be risky once a unique or a couple field trial is a random sample and does not represent all possible outcomes of a random effect, leading to low reliable results (Gaynor et al., 2021). Moreover, the response to selection using empirical data is only valid for the current generation (Falconer & Mackay, 2009) and cannot be extrapolated to future breeding cycles.

Core Ideas

- Building training sets using historical data outperforms the "test-and-shelf" approach.
- Genetic optimization algorithms to build training sets have no clear benefit in the long term.
- Introducing external elite germplasm reduces the genomic prediction accuracy.

2 | MATERIALS AND METHODS

Our study compared different strategies to compose GS TSs, the usefulness of genetic algorithms for "trimming" or "optimizing" the TS genetic variability, and the effect of introducing elite germplasm in a GS-based breeding program. For that, we consider rice as a model for self-pollinated crops and the advent of stochastic simulations via *AlphaSimR* (Gaynor et al., 2021), following the main steps in this type of study:

2.1 | Historical population and genetic parameters

The historical rice founder population was simulated as 3000 unique diploid inbred individuals, with 12 chromosome pairs each, using a Markovian Coalescent Simulator (Chen et al., 2009), considering a "GENERIC" species. The number of aggregating segments was defined based on the genome size (cM) described by L. Li et al. (2008). The "GENERIC" option allows the user to define specific genetic/genomic features to represent the species in the study as much as possible.

The target of the simulation was a quantitative trait, such as grain yield. The trait was composed of 30 quantitative trait nucleotides (QTNs) per chromosome, totaling 360 QTNs. A simulated single nucleotide polymorphism (SNP) chip with 83 SNPs per chromosome was used for genotyping, totaling 996 SNPs; SNP and QTN sites were not allowed to overlap. The additive, dominance, and average degree of dominance parameters were defined based on L. Li et al. (2008). Each QTN was assigned additive and dominance effects. Total genetic values for each genotype were obtained by summing all additive and dominance effects times the appropriately scaled genotype dosage for all QTNs; for details, see Gaynor et al. (2021). Additive effects (a) were sampled from a gamma distribution with scale and shape parameters equal to 1 and randomly assigned for each QTN. Similarly, dominance effects (d) for each QTN were computed by multiplying the

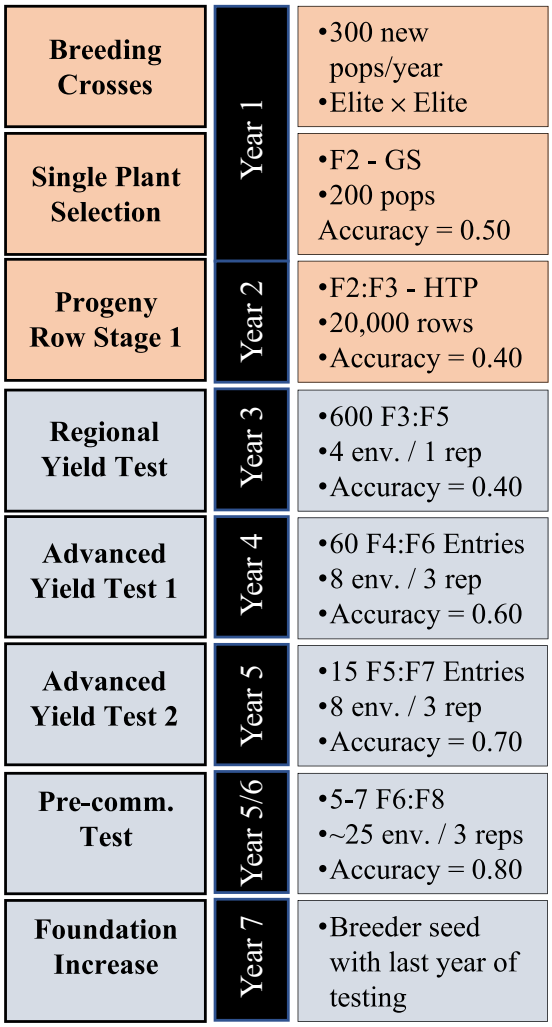

FIGURE 1 LSU Rice Breeding Scheme, population sizes, and the estimated selection accuracy per stage. GS, genomic selection.

absolute value of its additive effect (a_i) by locus-specific dominance degree (δ_i). Dominance degrees were sampled from a Gaussian distribution with $\delta_i \sim N(\mu_\delta, \sigma^2_\delta)$, where μ_δ is the average dominance degree equal to 0.22 and σ^2_δ is the variance of the dominance degrees equal to 0.125. Therefore, there is at least a 26% chance that the delta will be negative (bi-directional dominance deviations) and a 1% chance that it will exceed the unit (overdominance).

The initial mean of the quantitative trait was 0, and its initial total genetic variance was 1. Phenotypic values of each individual were obtained by adding a random error sampled from a Gaussian distribution to its true total genetic value such that initial broad-sense heritability was set according to the accuracy selection we observed in the Louisiana State University breeding program (Figure 1). Also, the heritability changed over the breeding cycles as genetic variances changed. This study did not consider epistasis or mutations, although it may contribute to heterosis or create genetic variability in rice (Huang et al., 2016).

2.2 | Base population, burn-in phase, and the first GS TS

In order to obtain our base population, we selected 60 individuals based on their superior phenotypic values from 3000 lines of the historical population. As a starting point, we considered a traditional representative program as a 5-year rice breeding program (from cross to cross), without GS or high-throughput phenotyping in the first breeding cycles, in other words, the previous version of our current breeding program (Figure 1). Overall, the breeding scheme represents an adaptation of the Pedigree method (Breseghello & Coelho, 2013). Based on that, we simulated five selection cycles totaling 25 years of breeding in the burn-in stage. In each cycle, 60 parental lines were crossed to generate 160 F_1 plants, which were selfed to produce 150 F_2 plants from each cross. After five breeding cycles, we obtained the base population to evaluate the downstream scenarios of this study.

Regarding the GS, the initial TS was composed of 1152 inbred lines from 30 crosses between 60 individuals (parents), with nearly 40 plants per cross from the base population after the burn-in stage. The marker effects were predicted using the ridge-regression best linear unbiased prediction (Endelman, 2011) according to the equation below:

$$\boldsymbol{y} = 1\mu + \boldsymbol{Z_u u} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is the vector of individual phenotypic values from the TS; μ is the mean (intercept); $\boldsymbol{u}$ is the vector of marker effects, where $\boldsymbol{u} \sim N(0, \boldsymbol{I}\sigma^2_u)$; and $\boldsymbol{\varepsilon}$ is the vector of random residuals. 1 is the vector of ones and $\boldsymbol{Z_u}$ is the incidence matrix of TS genotypes for m markers. $\boldsymbol{Z_u}$ is coded as 1 for homozygous A_1A_1, −1 for homozygous A_2A_2, and 0 for heterozygous A_1A_2.

To perform the GS, the genomic estimated breeding value (GEBV) was estimated using the following equation: GEBV = $\boldsymbol{Mu}$, where $\boldsymbol{M}$ is the incidence matrix of selection candidate genotypes and $\boldsymbol{u}$ is the vector of predicted marker effects. Overall, the goal in all breeding scenarios was to deploy GS in F_2 plants.

2.3 | Breeding scenarios simulated

2.3.1 | Approaches to build the TSs

We compare two main approaches to collect phenotypic and genotypic data and compose the TS, named "historical" and "test and shelf." In the former (Figure 2a), predictions rely on data accumulated in the last three breeding generations (representing more than three calendar years). Every year, new data come from the preliminary yield trials (PYT), which represent the $F_{3:4}$ breeding stage, to update the TS. For that, we adopted the grandparents, parents, offspring (GPO) strategy

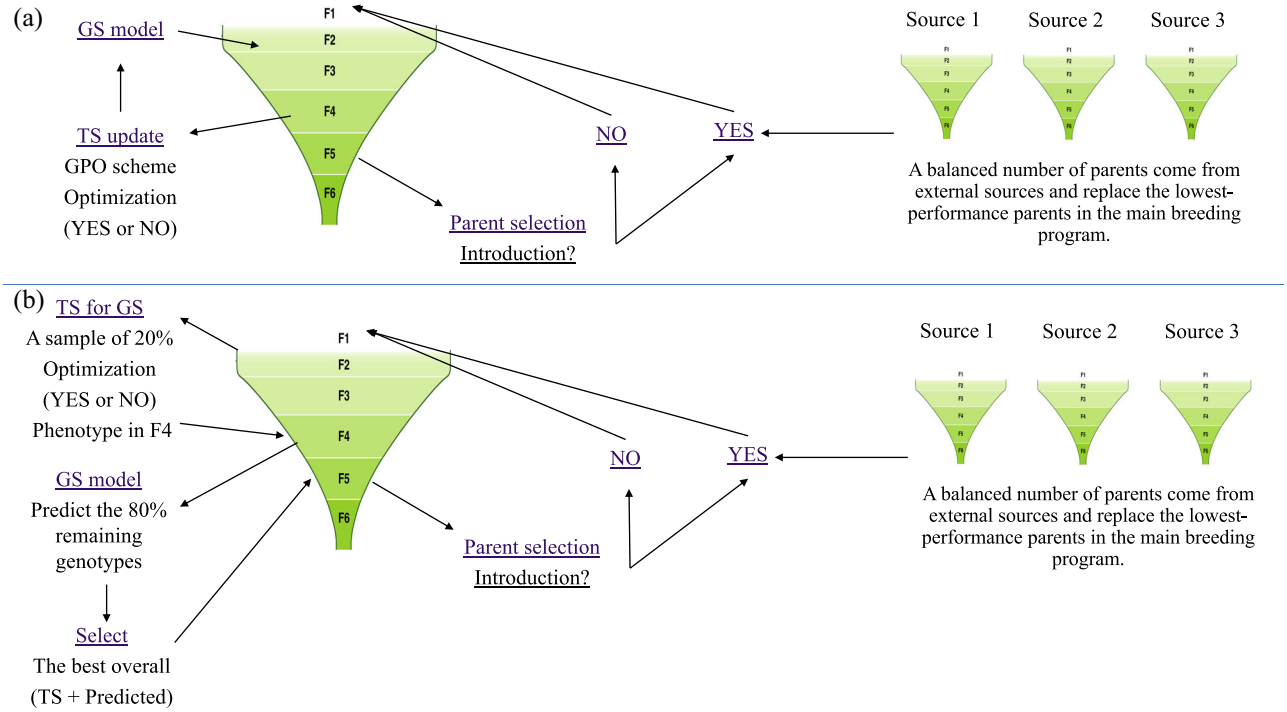

FIGURE 2 Training set schemes ([a] historical data and [b]"test and shelf"), the use of genetic optimization algorithms, and parent introgression rates per breeding cycle. GPO, grandparent, parent, offspring; GS, genomic selection; TS, training set.

to keep the accuracy at reasonable levels for more breeding cycles (Sabadin et al., 2022). In short, this strategy considers only the last three breeding cycles to compose the TS (GPO). Therefore, we added the newest data for every breeding cycle, removed the oldest one, and maintained only the last three breeding generations in the TS. The best individuals selected in the F_2 generation are advanced. Then, after a round of phenotypic selection in F_5, the best individuals are selected as parents.

In the second approach (Figure 2a), predictions rely on a proportion of the progeny (20%) that is advanced in order to work as a TS for the remaining part (80%) that will "wait" in the cold room. Therefore, there is a brand-new TS every year, with the maximum relationship between it and the prediction set. On the other hand, the year effect highly affects the phenotypic information. As described in the former method, the new data are updated to update the TS from the PYT, which represents the $F_{3:4}$ breeding stage. The best individuals are selected from the whole dataset, TS, and predicted set together. Then, after a round of phenotypic selection in F_5, the best individuals are selected as parents.

2.3.2 | Optimizing TS (OTS) using genetic algorithms

In order to define the OTS, we used the method proposed by Akdemir et al. (2015), with a predefined population size. For "test and shelf" (Figure 2b), we used the algorithm to select 20% of the individuals genetically representing the whole population. In its turn, in the "historical" scenario, the algorithms were used to "trim" the historical dataset, keeping 99% genetic variability useful to predict the newest F_2 population (Figure 2a).

In this methodology, the selection of lines requires only genotypic information on the individuals present in a group of candidates and the target. Subsequently, based on this, a genetic algorithm makes an approximation of the prediction error variance (PEV) using principal components via the marker matrix and selects determined hybrids that will establish the OTS. Assuming P denotes the first 50 principal components corresponding to the lines genomic additive matrix (VanRaden, 2008), the PEV for predicting the genotypes that are not included in the TS is approximated by $\text{PEV} = \text{tr}(P_{\text{Test}}(P'_{\text{Train}}P_{\text{Train}} + \lambda I)^{-1}P_{\text{Test}})$, where P_{Test} and P_{Train} correspond to the principal components of test (lines that are not considered for training) and training individuals, λ is a small positive real number (1E–5, in this case), and tr is the trace operator that takes the sum of diagonals of a square matrix. The algorithm for establishing OTS was implemented via the STPGA R package, considering 100 interactions per scenario (Akdemir, 2017). Subsequently, we obtain the OTS and identify the lines to compose them. As baselines, we conducted the same breeding strategies by sampling 20% of the lines by chance to build a TS for "test

and shelf" or considering the historical data as a whole (no trimming) for the other scenario.

2.3.3 | Introducing elite germplasm

In this factor, we consider four possible rates of elite germplasm introduction: 0%, 10%, 20%, or 30%, where 0% means a closed breeding program; in other words, only the parents from the breeding programs will recycle the crossing block. In the other cases, the introductions come from three other breeding programs in a one-way migration and balanced proportion. This scenario wants to mimic situations such as germplasm acquisition, the breeding program working as a hub, or breeding consortiums, but with planned introduction rates per breeding generation. For instance, considering that the crossing block is composed of 60 parents and the aim is to introduce 10% of external germplasm, the 54 best parents will be kept, and six parents, being the two best parents from each one of the three external sources, will compose the set of new parents.

2.4 | Comparing breeding schemes

Considering that all compared methods used almost the same framework, there were no significant differences in cycle length among them. Therefore, we measured the average true breeding value of the lines, the true genetic value of the best line, the true additive genetic variance, the prediction accuracy, and the divergence between the main breeding programs and the sources of elite germplasm over the breeding cycles. The prediction accuracy was calculated using Pearson's correlation between true and estimated breeding values. In turn, the divergence was assessed by the fixation index (Fst) (Luo et al., 2019). Each strategy was simulated for ten breeding cycles and replicated 100 times within a single population using the *AlphaSimR* package (Gaynor et al., 2021). All scripts necessary to run the analysis were provided as the Supporting information.

3 | RESULTS

3.1 | Approaches to build the TSs

Two primary approaches for composing TSs were compared: the "test-and-shelf" method and the "historical" method (Figure 3). On average, the historical strategy yielded higher population means and led to the best line's highest performances throughout 10 breeding cycles. Notably, in the initial breeding cycle, there was a substantial increase in the performance of the best line; however, this incremental gain decreased gradually from the second cycle onward until the curve approached a plateau. Concurrently, genetic variability (additive variance) exhibited a significant decrease until the second breeding cycle, followed by a slower decline, eventually reaching a value close to zero by the 10th cycle.

Predictive accuracy mirrored this trend, showing a sharp decrease until the second cycle but subsequently rebounding to attain intermediate values compared to the test-and-shelf approach. Notably, the latter method differed markedly from the historical approach regarding predictive accuracy, maintaining high levels throughout the 10 years of the GS-based breeding program. While its additive variance initially displayed higher values in the first cycles before gradually declining, it remained higher than the "historical" approach.

An intriguing finding is that despite the test-and-shelf method demonstrating greater preservation of long-term variability and predictive accuracy, it did not efficiently translate these resources into genetic gain. Additionally, it is essential to highlight that the graph represents an average of the performance of both methodologies, with the impacts of different introgression rates and TS optimization intertwined in these values.

3.2 | Optimizing the TS composition using genetic algorithms

The optimization algorithm was applied differently in composing the TSs for the test-and-shelf and historical approaches. In the former, optimization was employed to select 20% of individuals representing population diversity. In contrast, the latter involved trimming the historical dataset while maintaining 99% of the population's genetic variability.

A notable finding is that there is no discernible advantage in utilizing TS optimization with the test-and-shelf approach (Figure 4). Furthermore, in the historical method, employing the optimization algorithm proved disadvantageous for maintaining prediction accuracy and additive variance in the long term. However, it did not confer any advantage when analyzing population mean and best line performance.

3.3 | Introducing elite germplasm

The last factor investigated was the percentage of elite germplasm introduced from external breeding programs, ranging from 0% to 30% (Figure 5). Concerning population mean and best line performance, both the historical and test-and-shelf approaches exhibited similar increasing patterns across the breeding cycles. However, the historical method remained unaffected by varying introgression rates, whereas the test-and-shelf approach displayed its poorest performance when operating as a closed breeding program (0% introgression).

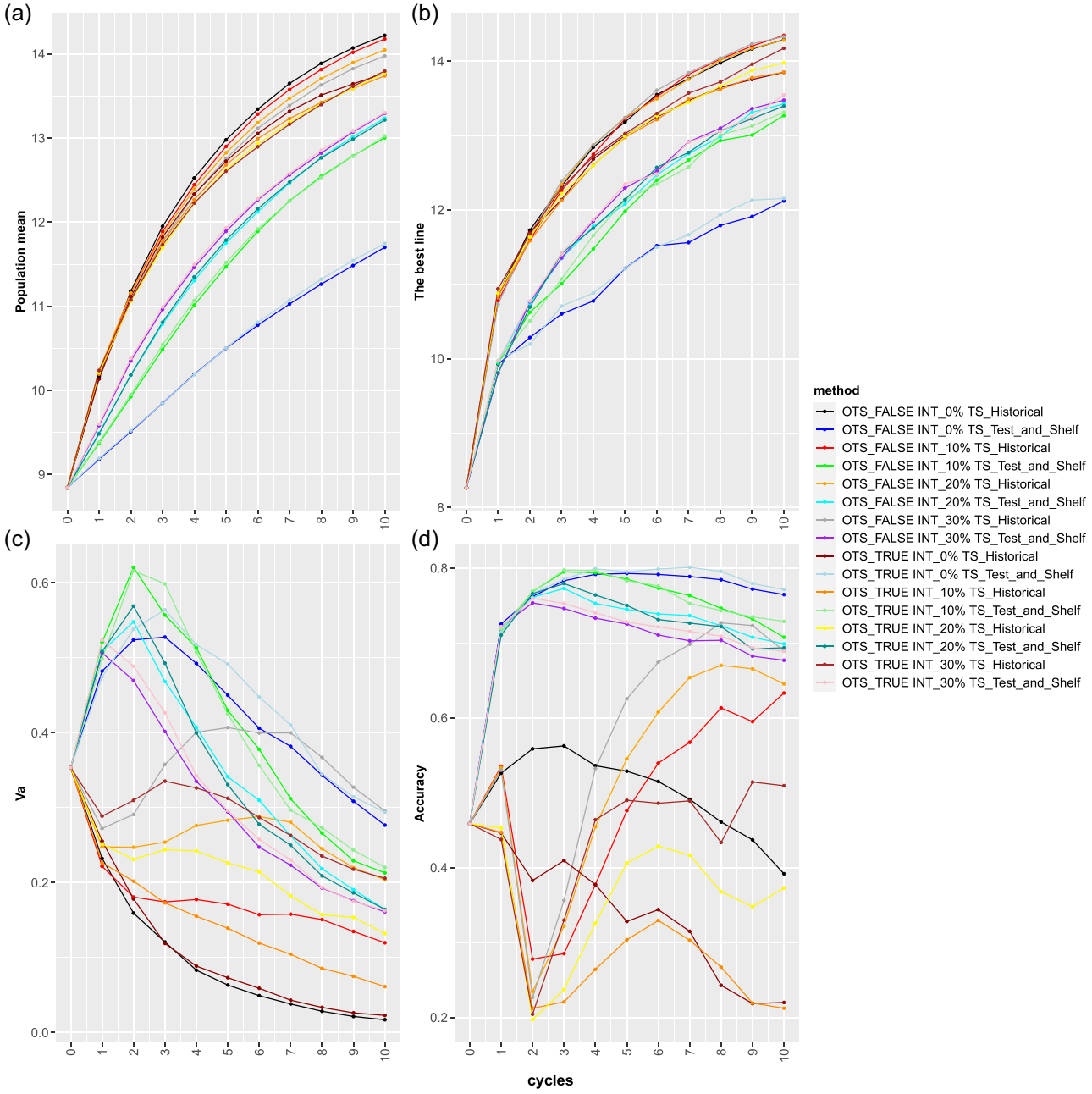

FIGURE 3 Average population performance means (a), the best line (b), additive variance (c), and prediction accuracy (d) over 10 breeding cycles considering different scenarios of training set composition and the use of genetic optimization algorithms, regardless of the rates of external elite germplasm introduction. OTS, optimized training set.

Regarding additive variance and prediction accuracy, the two approaches reacted divergently to different introgression percentages. The historical approach demonstrated higher values for these parameters under an open system program (with more than 10% introgressions), while the test-and-shelf approach yielded the highest values in a closed breeding program scenario. Additionally, the historical approach experienced a decline in accuracy in the second cycle. Still, over time and with consistent material introgression, accuracy levels were gradually recovered.

3.4 | Comparing breeding schemes

The final step in comparing the long-term effects of breeding schemes involved analyzing the performance of all tested parameter combinations, including the approach used to compose the training population, the utilization of the TS optimization algorithm, and varying introgression rates (Figure 6). Overall, the historical approach (represented by warm colors) demonstrated superior performance in terms of selection gain. However, this gain was accompanied by a

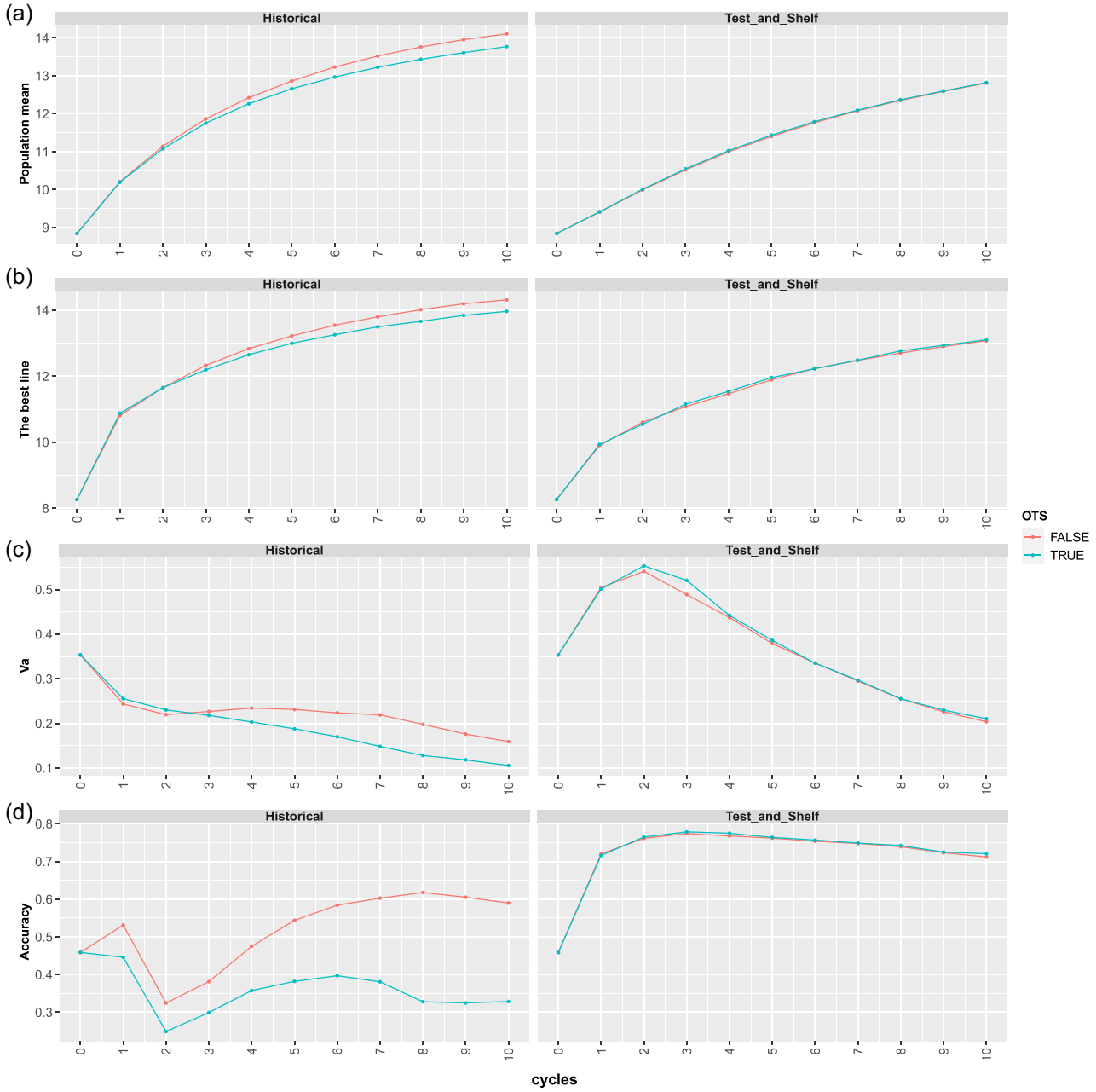

FIGURE 4 Average population performance means (a), the best line (b), additive variance (c), and prediction accuracy (d) over 10 breeding cycles considering different scenarios of training set composition and use of genetic optimization algorithms, regardless of external elite germplasm introduction. OTS, optimized training set.

genetic variability and accuracy loss over the years. Notably, certain parameter combinations within this approach resulted in intermediate losses in genetic variability and predictive accuracy, such as the cases with 30% and 10% introgression without TS optimization (depicted by gray and orange lines).

Conversely, the test-and-shelf approach (represented by cold colors) exhibited the best prediction accuracies and preservation of genetic variability, albeit without translating these advantages into population genetic means. For instance, combinations like the dark blue and light grayish-blue lines (0% introgression, with and without TS optimization) showcased the highest values of predictive accuracy and preservation of genetic variability in the long term yet achieved the lowest genetic gains. However, similar to the historical approach, test and shelf also produced intermediate results, as evidenced by combinations like the pink and purple lines (30% introgression of materials, with and without TS optimization). While these combinations yielded lower genetic gains than all historical approach combinations, they closely approached them in performance.

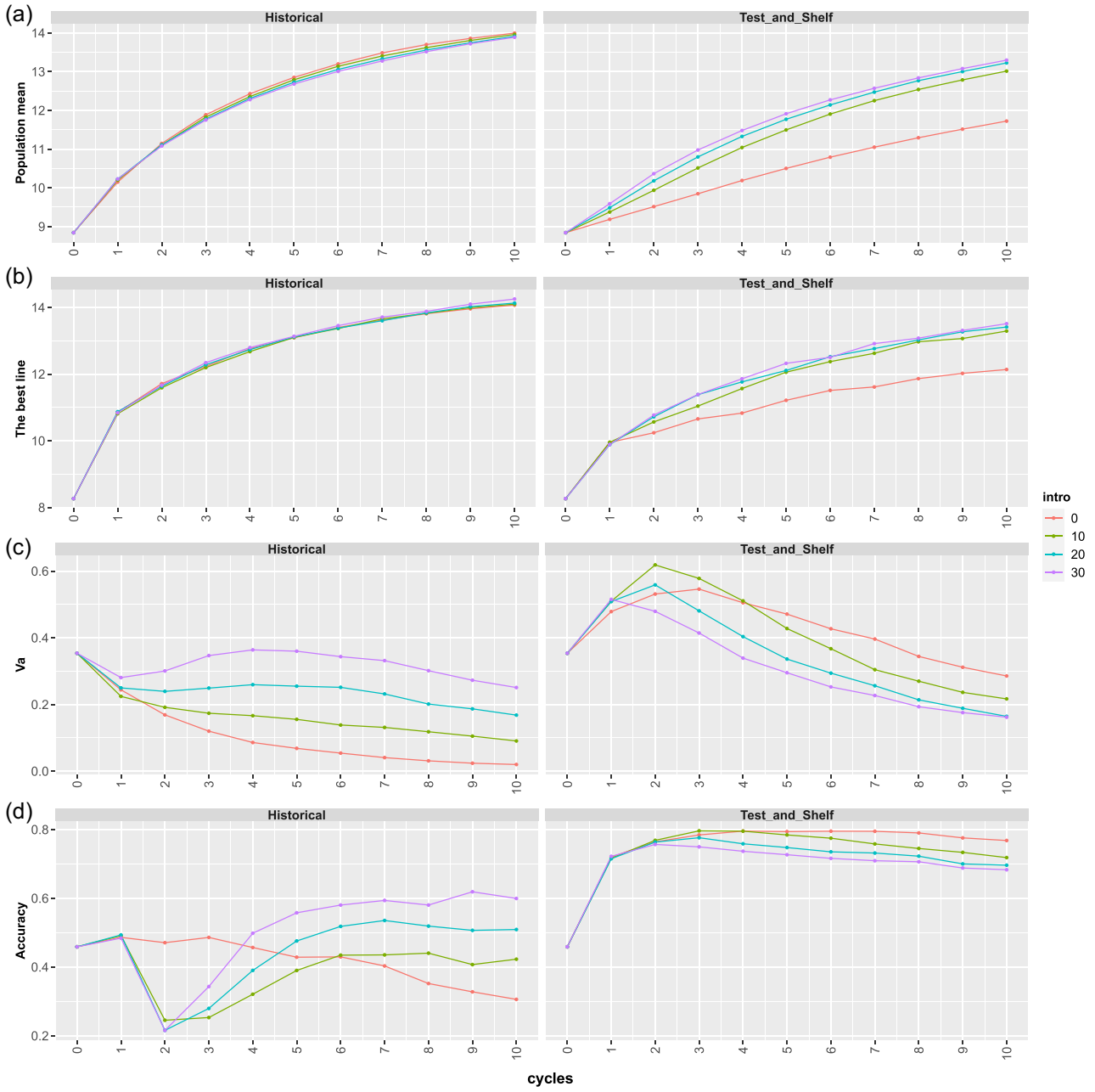

FIGURE 5 Average population performance means (a), the best line (b), additive variance (c), and prediction accuracy (d) over 10 breeding cycles considering different scenarios of training set composition and rates of external elite germplasm introduction (Intro), regardless of the use of genetic optimization algorithms.

3.5 | The one-way introduction effect on the genetic distance between breeding programs

Finally, the Fst between the main breeding program and the other three external programs was calculated in order to understand their genetic relationship throughout the years (Figure 7). In the first cycles, the relationship between the main program and the other three external ones was high, with a Fst close to zero. With no introgression, the closed system breeding program distanced itself from external programs as the years passed. Meanwhile, programs with annual material introduction were able to maintain a high relationship with external breeding programs, maintaining a low index value during the 10 cycles considered. As expected, the higher the percentage introduced, the smaller the increments in Fst over breeding cycles. These results have been observed in other empirical studies, where germplasm exchange between breeding programs has resulted in low genetic differentiation and high genetic diversity within germplasm collections (Delfini et al., 2021; Tsindi et al., 2023). The Fst value still slightly increases over cycles, primarily due to artificial selection and genetic drift in each breeding program and also because the introductions were made in just one direction, in other words, from external sources to the main breeding program, but

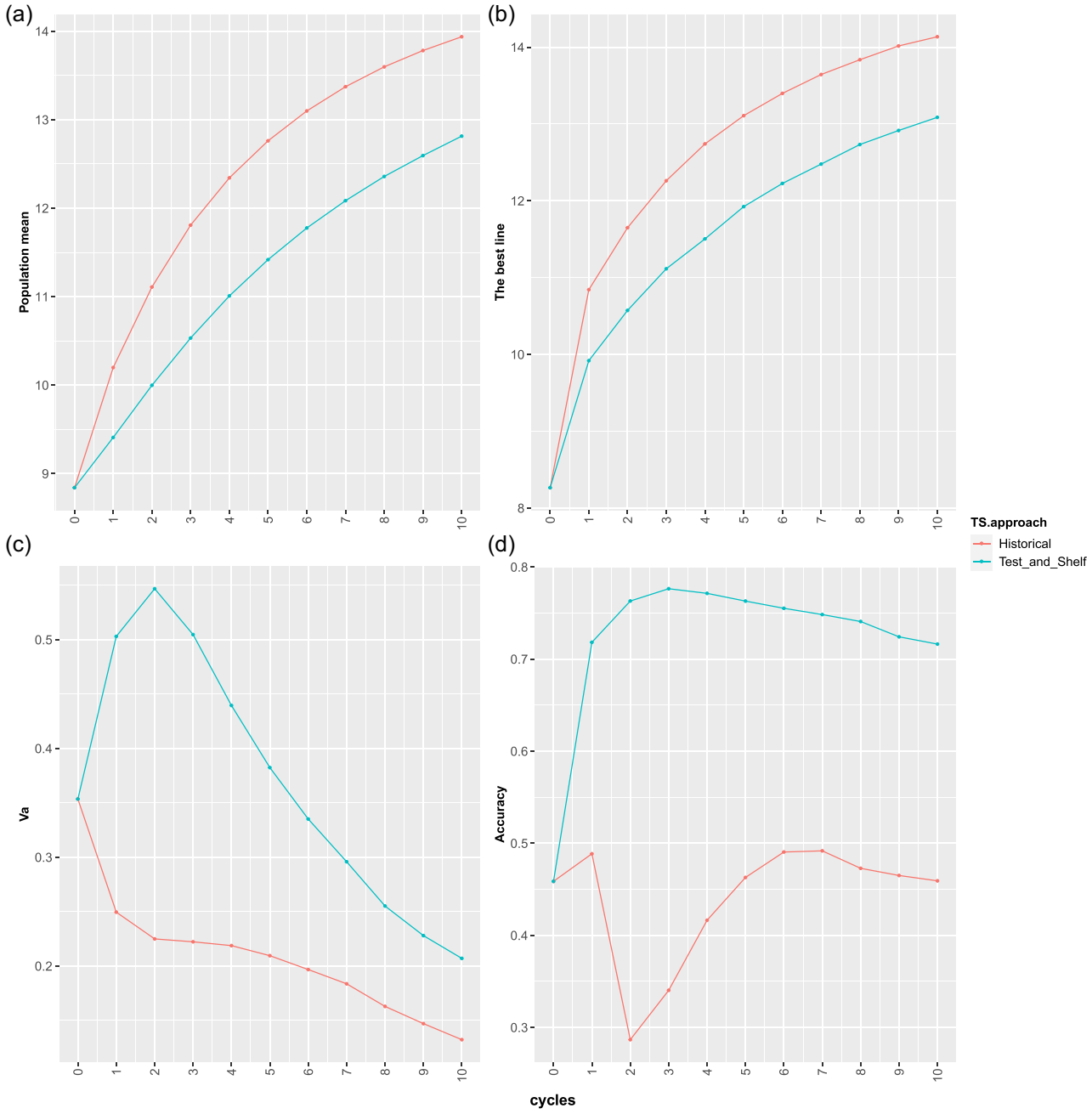

FIGURE 6 Population performance means (a), the best line (b), additive variance (c), and prediction accuracy (d) over 10 breeding cycles considering different scenarios of training set composition, the use of genetic optimization algorithms, and external elite germplasm introduction. TS, training set.

the opposite was not true. These factors naturally shape the genetic structure among populations through allele frequency changes and favorable alleles accumulation across breeding cycles (Kolawole et al., 2017).

4 | DISCUSSION

Plant breeding is a complex, long-term process requiring significant time and resource investment. Decisions within a breeding program must be made with utmost caution, as their short-and long-term impacts can be critical to the program's success. With an ever-increasing amount of data, there is a gap between developing plant breeding practices and applying new information that could optimize processes within the breeding programs. This gap may result from the time needed to develop breeding program practices or the risk of negative impacts that a wrong decision could have on maintaining the program (X. Li et al. 2012). Simulations offer breeders the opportunity to explore a wide range of conditions of interest,

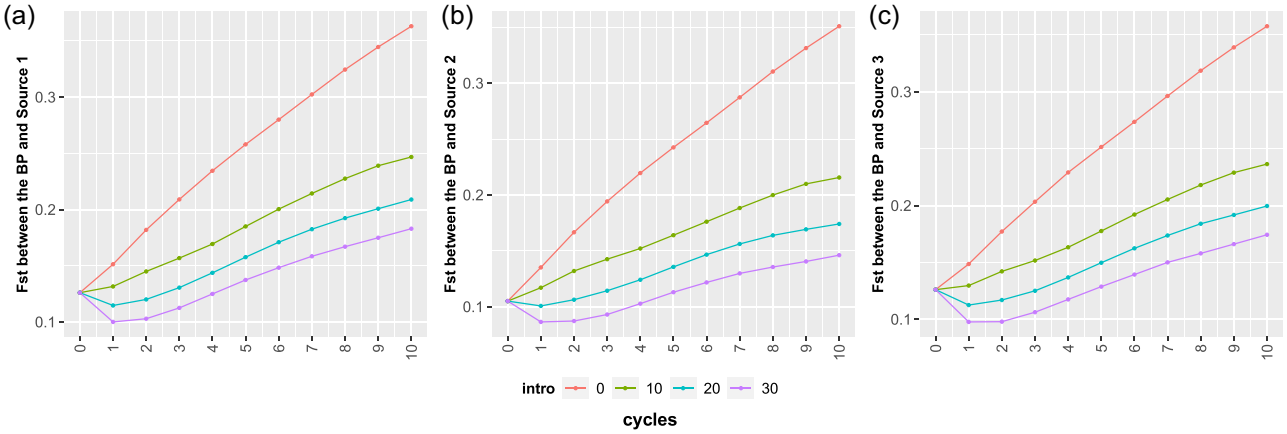

FIGURE 7 Fixation index (Fst) between the main program and the three external sources, over 10 breeding cycles, considering different rates of external elite germplasm introduction (Intro).

aiding in developing and testing strategies to maximize genetic gain, preserve genetic diversity, optimize operational costs, improve parental crosses, and introduce trait variation from external sources (Y. Li, Shi, et al. 2022). Therefore, stochastic simulation emerges as a valuable tool for addressing this challenge, facilitating the cost-effective design and optimization of breeding programs (Gaynor et al., 2021). For instance, they enable monitoring the impact of strategy adoption on various aspects over dozens of breeding cycles, such as maintaining genetic diversity based on the selection scheme (Allier et al., 2020), training population designs and genotyping strategies (Hickey et al., 2014), the number of parents, the number of hybrids, tester updates, and the genomic prediction of hybrid performance (Fritsche-Neto et al., 2024), among others. In contrast, empirical studies often depend on assumptions from static data collected at one or a few points in time to make inferences about the past and/or present.

Researchers utilize genetic algorithms and methodologies to optimize TS composition in GS studies. For instance, Fritsche-Neto et al. (2018) compared different sample selection methods to random sampling with biparental populations. Lorenz and Smith (2015) combine data from multiple related and/or unrelated individuals, while Rincent et al. (2012) focus on incorporating diverse populations in the TS. This last one, with criteria from the mixed model equations, the coefficient of determination, and the PEV. Additionally, Isidro et al. (2015) introduced stratified sampling and stratified coefficient of determination as alternative algorithms aimed at enhancing the optimization of TSs, particularly under the influence of population structure effects, with stratified methods exhibiting superior performance, particularly when population structure effects were prominent. By doing so, breeders can fully leverage GS to accelerate genetic gain and enhance the efficiency of their breeding programs (Fernández-González et al., 2023).

Concerning the TS compositions, overall, those composed of historical data provided higher genetic gains than test and shelf. Therefore, the advantage of having more years of data, even without the maximum connectivity, leads to a more accurate estimation of breeding values. On the other hand, the test-and-shelf scenarios were much more stable in terms of accuracy, even when elite germplasm introgressions were considered. The reason resides in the fact that a brand-new TS is composed every year, using a random part of the new population, ensuring the highest level of connectivity between the training and testing sets and representing all the "new alleles" and haplotypes in their current frequencies and linkage disequilibrium (LD) patterns but not necessarily reflecting in genetic gain.

In this context, some studies analyzed the optimization of long-term breeding schemes, aiming to balance increased selection gains with the preservation of genetic variability (Gorjanc et al., 2018; Obšteter et al., 2019; Pocrnic et al., 2023; Sabadin et al., 2022; Wientjes et al., 2022). However, we present an intriguing result regarding the efficiency of "test and shelf" in transforming genetic variability into selection gain. The two breeding schemes we used started from the same base population after the burn-in stage and had identical selection intensities, cycle lengths, and heritabilities. However, only the "test-and-shelf" approach maintained consistent rates of selection gain and lower population means across the cycles. For instance, the dark blue and light grayish-blue lines (0% introgression, "test-and-shelf" approach) appear far from reaching the population mean curve plateau, representing significant preservation of additive variance. Conversely, the black and dark red lines (0% introgression, "historical" approach) achieved the highest population means but quickly exhausted the population's additive variance. This observation might lead to the misconception that the "test-and-shelf" approach has lower predictive accuracy but provides the highest. A plausible explanation is that the "test-and-shelf"

training population, although accurately representing the haplotypes in the test population, includes only genotypes selected from the current cycle with residues from only 1 year. This environmental residue might erroneously exclude some important haplotypes for the trait, thus reducing selection gain. This could also explain why open breeding schemes using the "test-and-shelf" approach resulted in an increased population mean despite a decrease in accuracy relative to their test population. In contrast, the "historical" approach includes the last three generations in the training population, featuring haplotypes from the best lines over the last three annual cycles. In addition, the test-and-shelf approach does not consider multi-year genomic information; in other words, the allelic frequency across years might change drastically, depending on the TS of each cycle, keeping high genetic variance and accuracy. Still, it might lower the mean when compared with the multi-year approach because, in each cycle, a group of different haplotypes will be recombined and kept to improve its frequency in the population, reducing the population mean. Although the historical TS shows lower accuracy with the current cycle's population, it provides a more precise representation of the important haplotypes for the target trait.

Observing the additive variance trends of both approaches demonstrates further evidence that important haplotypes are being excluded from the training population in the "test-and-shelf" approach (Figure 4). GS typically results in a significant drop in additive variance during the first cycles, known as the Bulmer effect (Bulmer, 1971), as observed with the "historical" approach. However, the "test and shelf" showed an opposite trend, probably because the approach "kept" the allele frequencies almost constant, or the most probable explanation is that it increased the frequency of minor alleles, consequently "compensating" the losses due to drift and selection. It is well-known that the highest levels of additive variance occur when loci alleles have a frequency of 0.5 in populations of self-pollinated crops (Falconer & Mackay, 2009). Thus, the "test-and-shelf" approach initially led to a more balanced allele frequency, then experienced a decline as selection cycles progressed, indicating a change in alleles previously selected within this breeding scheme. Again, this change suggests that considering only 1 year in the training population directs the selection toward just one haplotype group that is not necessarily the most important for the trait.

The deployment of optimization algorithms aimed at trimming datasets to enhance prediction accuracy often fails to yield substantial gains in population improvement or the preservation of genetic variability and accuracy over time. This limitation stems from the iterative fine-tuning process inherent in optimization, which may inadvertently accumulate "constraints" or lead to overfitting across breeding cycles, particularly within the context of the historical approach used to build the TS (Neyhart et al., 2017). For instance, consider a scenario where an optimization algorithm is employed to select a subset of individuals for the TS based on their genetic markers. Initially, this may result in improved prediction accuracy for the target traits. However, the algorithm's emphasis on selecting individuals with known favorable alleles over successive breeding cycles may inadvertently limit the genetic diversity within the TS (Muleta et al., 2019). Consequently, the TS becomes increasingly specialized toward known alleles (Sabadin et al., 2022), potentially overlooking rare alleles or novel genetic combinations that could contribute to transgressive phenotypes (Isidro et al., 2015). In practice, this limitation can be observed in a breeding program that utilizes the historical approach, where the TS is constructed based on past breeding data. Initially, the optimization algorithm may effectively select individuals with favorable traits for inclusion in the TS. However, as the breeding program progresses, the algorithm's tendency to prioritize known alleles may reduce the representation of genetic diversity, hindering its ability to predict the performance of novel genotypes (DoVale et al., 2022) or identify transgressive combinations. Furthermore, the high computational demand associated with optimization algorithms poses a practical challenge, particularly when working with large populations. Addressing these challenges will require a nuanced approach that balances the benefits of optimization with the need to preserve genetic variability and adaptability in breeding populations.

Finally, regarding the elite germplasm introgression, the "historical" approach showed to be very stable in terms of introduction rates concerning population improvement, with a significant drop in accuracy in the first class, but over the cycles, considering that the introduction will content and from the same source, recovered the accuracy, balancing the increasing of genetic variability available and losses in prediction ability. On the other hand, the "test-and-shelf" approach was shown to be well designed in cases of germplasm introduction, keeping the prediction accuracies much more stable.

Similar to our results, studies involving multibreed reference populations in cattle breeding have also observed a drop in prediction accuracy when predictions are made across breeds. The breeds dominate SNP effects in greater proportion within the reference population, and the prediction model captures the effects of SNPs that exhibit the same LD pattern with the QTLs across all breeds or only in the largest population. This can lead the model to predict a non-existing SNP effect in other breeds, resulting in a loss of SNP prediction ability (Karaman et al., 2021; van den Berg et al., 2016).

We also observed that introducing germplasm using the "historical" approach led to a quicker recovery and overall improvement in prediction accuracy as the introduction proportion increased (Figure 5d). This might allow the model to equalize the SNP effects across the germplasm

better. Conversely, the "test-and-shelf" approach exhibited an opposite pattern where increased introduction proportion decreased prediction accuracy. The presence of different LD patterns between SNPs and QTLs could have introduced "noise" into the closed system and reduced the prediction accuracy. Meanwhile, programs with annual material introduction maintained a high relationship with external breeding programs, keeping a low index value during the 10 cycles considered. As expected, the higher the percentage introduced, the smaller the increments in Fst over breeding cycles.

Other studies via stochastic simulations have studied germplasm exchange and its consequences. For instance, Yabe et al. (2016) used the Island model to understand how small-breeding programs can exchange germplasm and improve their relationship to increase the GS TS sizes and accuracies. The proposed scheme better maintains genetic improvement in later generations than the other GS methods, suggesting that the Island-model GS can utilize genetic variation in breeding and retain alleles with small effects in the breeding population. Another interesting study was performed by Technow et al. (2021) to understand how the genetic gains happened, considering the structure of commercial plant breeding programs, particularly in major crops like maize, is characterized by a large degree of decentralization with the exchange of successful germplasm within but not across companies.

Besides the interesting findings, the abovementioned ones did not cover the aspects of the present study. Of course, a limitation of this manuscript is that we considered a one-way introduction; in other words, a main breeding program brings new variability from others, but the opposite is not true, which is a potential next hypothesis. Moreover, we did not consider epistasis; besides, it may impact the genetic variability, genotypic values, and the response to selection. The bottleneck to include epistasis in simulations is to quantify the real importance of this genetic component in commercial breeding programs. Finally, we considered just one algorithm of optimization, which may not be the best in terms of stability and response across all crops and traits (Fernández-González et al., 2023).

Conclusively, considering the practical consequences of TS compositions in breeding programs, the test and shelf seems more stable regarding accuracy in dealing with introductions despite the origin, frequency, and rate. Therefore, it may be useful in programs where the introductions come from different sources over the years; in other words, they are not programmed. Conversely, using historical data, if the introductions come from the same source over the cycles, this negative effect is reduced as long as the cycles of introductions and this approach are the best. Furthermore, it may support public breeding programs in establishing consortiums or networks of collaborations, where the exchange of germplasm will occur at a predefined rate and flow. In either case, the use of algorithms of optimization to trim the genetic variability does not bring a substantial advantage in the medium to long term. It may be more useful in budling the first TS from big and heterogeneous historical data. Ultimately, it is important to highlight that in this study, we used population sizes that were more similar to those observed in public breeding programs and the absence of genotype × year due to package limitations. Therefore, the advantages or disadvantages of the test and shelf may vary depending on the population size, the crop, and the trait considered as long as the year effect causes crossover interactions. Consequently, more studies in this context are needed.

AUTHOR CONTRIBUTIONS

Roberto Fritsche-Neto: Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; supervision; visualization; writing—original draft. **Rafael Massahiro Yassue**: Conceptualization; investigation; methodology; writing—original draft. **Allison Vieira da Silva**: Conceptualization; writing—original draft. **Melina Prado**: Investigation; writing—original draft. **Júlio César DoVale**: Conceptualization; investigation; writing—original draft.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets and scripts used for this study can be found in the supplementary material.

ORCID

Roberto Fritsche-Neto https://orcid.org/0000-0003-4310-0047
Melina Prado https://orcid.org/0000-0001-5926-1617

REFERENCES

Akdemir, D. (2017). STPGA: Selection of training populations with a genetic algorithm. *BioRxiv*, https://doi.org/10.1101/111989

Akdemir, D., & Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports*, *9*, Article 1446. https://doi.org/10.1038/s41598-018-38081-6

Akdemir, D., Sanchez, J. I., & Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*, Article 38. https://doi.org/10.1186/s12711-015-0116-6

Alemu, A., Åstrand, J., Montesinos-López, O. A., Isidro Y Sánchez, J., Fernández-Gónzalez, J., Tadesse, W., Vetukuri, R. R., Carlsson, A. S., Ceplitis, A., Crossa, J., Ortiz, R., & Chawade, A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant.*, *17*, 552–578. https://doi.org/10.1016/j.molp.2024.03.007

Allier, A., Teyssèdre, S., Lehermeier, C., Moreau, L., & Charcosset, A. (2020). Optimized breeding strategies to harness genetic resources with different performance levels. *BMC Genomics*, *21*, Article 349. https://doi.org/10.1186/s12864-020-6756-0

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, *34*, 20–25. https://doi.org/10.2135/cropsci1994.0011183X003400010003x

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 Genetics*, *18*, Article 51. https://doi.org/10.1186/s12863-017-0512-8

Berro, I., Lado, B., Nalin, R. S., Quincke, M., & Gutiérrez, L. (2019). Training population optimization for genomic selection. *The Plant Genome*, *12*, 190028. https://doi.org/10.3835/plantgenome2019.04.0028

Beyene, Y., Gowda, M., Pérez-Rodríguez, P., Olsen, M., Robbins, K. R., Burgueño, J., Prasanna, B. M., & Crossa, J. (2021). Application of genomic selection at the early stage of breeding pipeline in tropical maize. *Frontiers in Plant Science*, *12*, 685488. https://doi.org/10.3389/fpls.2021.685488

Boyles, R. E., Ballén-Taborda, C., Brown-Guedira, G., Costa, J., Cowger, C., Dewitt, N., Griffey, C. A., Harrison, S. A., Ibrahim, A., Johnson, J., Lyerly, J., Marshall, D. S., Mason, R. E., Mergoum, M., Murphy, J. P., Santantonio, N., Saripalli, G., Sutton, R., Tiwari, V., … Winn, Z. J. (2024). Approaching 25 years of progress towards *Fusarium* head blight resistance in southern soft red winter wheat (*Triticum aestivum* L.). *Plant Breeding*, *143*, 66–81. https://doi.org/10.1111/pbr.13137

Breseghello, F., & Coelho, A. S. G. (2013). Traditional and modern plant breeding methods with examples in rice (*Oryza sativa* L.). *Journal of Agricultural and Food Chemistry*, *61*, 8277–8286. https://doi.org/10.1021/jf305531j

Bulmer, M. G (1971). The effect of selection on genetic variability. *The American Naturalist*, *105*(943), 201–211. https://doi.org/10.1086/282718

Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Research*, *19*, 136–142. https://doi.org/10.1101/gr.083634.108

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., De Los Campos, G., Burgueño, J., Windhausen, V. S., Buckler, E., Jannink, J.-L., Lopez Cruz, M. A., & Babu, R. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 Genes|Genomes|Genetics*, *3*, 1903–1926. https://doi.org/10.1534/g3.113.008227

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, *22*, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Delfini, J., Moda-Cirino, V., Dos Santos Neto, J., Ruas, P. M., Sant'ana, G. C., Gepts, P., & Gonçalves, L. S. A. (2021). Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm. *Scientific Reports*, *11*, Article 2964. https://doi.org/10.1038/s41598-021-82437-4

de Freitas Mendonça, L., & Fritsche-Neto, R. (2020). The accuracy of different strategies for building training sets for genomic predictions in segregating soybean populations. *Crop Science*, *60*, 3115–3126. https://doi.org/10.1002/csc2.20267

DoVale, J. C., Carvalho, H. F., Sabadin, F., & Fritsche-Neto, R. (2022). Genotyping marker density and prediction models effects in long-term breeding schemes of cross-pollinated crops. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *135*, 4523–4539. https://doi.org/10.1007/s00122-022-04236-3

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*, 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

e Sousa, M. B., Galli, G., Lyra, D. H., Granato, Í. S. C., Matias, F. I., Alves, F. C., & Fritsche-Neto, R. (2019). Increasing accuracy and reducing costs of genomic prediction by marker selection. *Euphytica*, *215*, Article 18. https://doi.org/10.1007/s10681-019-2339-z

Falconer, D. S., & Mackay, T. (2009). *Introduction to quantitative genetics* (4th ed.). Pearson, Prentice Hall.

Fernández-González, J., Akdemir, D., & Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *136*, Article 30. https://doi.org/10.1007/s00122-023-04265-6

Fristche-Neto, R., Akdemir, D., & Jannink, J.-L. (2018). Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. *Theoretical and Applied Genetics*, *131*, 1153–1162. https://doi.org/10.1007/s00122-018-3068-8

Fritsche-Neto, R., Ali, J., De Asis, E. J., Allahgholipour, M., & Labroo, M. R. (2024). Improving hybrid rice breeding programs via stochastic simulations: Number of parents, number of hybrids, tester update, and genomic prediction of hybrid performance. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *137*, Article 3. https://doi.org/10.1007/s00122-023-04508-6

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: An R package for breeding program simulations. *G3 Genes|Genomes|Genetics*, *11*(2), jkaa017. https://doi.org/10.1093/g3journal/jkaa017

Gonzalez, M. Y., Zhao, Y., Jiang, Y., Stein, N., Habekuss, A., Reif, J. C., & Schulthess, A. W. (2021). Genomic prediction models trained with historical records enable populating the German ex situ genebank bio-digital resource center of barley (*Hordeum* sp.) with information on resistances to soilborne barley mosaic viruses. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *134*, 2181–2196. https://doi.org/10.1007/s00122-021-03815-0

Gorjanc, G., Gaynor, R. C., & Hickey, J. M (2018). Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics*, *131*(9), 1953–1966. https://doi.org/10.1007/s00122-018-3125-3

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*, 1–12. https://doi.org/10.2135/cropsci2008.08.0512

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., Grondona, M., Zambelli, A., Windhausen, V. S., Mathews, K., & Gorjanc, G. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Science*, *54*, 1476–1488. https://doi.org/10.2135/cropsci2013.03.0195

Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q. I., Zhan, Q., Zhao, Y., Li, W., Cheng, B., Xia, J., Chen, N., Huang, T., Zhang, L., Fan, D., Chen, J., Zhou, C., Lu, Y., Weng, Q., & Han, B. (2016). Genomic architecture of heterosis for yield traits in rice. *Nature*, *537*, 629–633. https://doi.org/10.1038/nature19760

Isidro, J., Jannink, J.-L., Akdemir, D., Poland, J., Heslot, N., & Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *128*, 145–158. https://doi.org/10.1007/s00122-014-2418-4

Isidro y Sánchez, J., & Akdemir, D. (2021). Training set optimization for sparse phenotyping in genomic selection: A conceptual overview. *Frontiers in Plant Science*, *12*, 715910. https://doi.org/10.3389/fpls.2021.715910

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & De Los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics [Theoretische Und Angewandte Genetik]*, *127*, 595–607. https://doi.org/10.1007/s00122-013-2243-1

Juma, R. U., Bartholomé, J., Thathapalli Prakash, P., Hussain, W., Platten, J. D., Lopena, V., Verdeprado, H., Murori, R., Ndayiragije, A., Katiyar, S. K., Islam, M. D. R., Biswas, P. S., Rutkoski, J. E., Arbelaez, J. D., Mbute, F. N., Miano, D. W., & Cobb, J. N. (2021). Identification of an elite core panel as a key breeding resource to accelerate the rate of genetic improvement for irrigated rice. *Rice*, *14*, Article 92. https://doi.org/10.1186/s12284-021-00533-5

Kaler, A. S., Purcell, L. C., Beissinger, T., & Gillman, J. D. (2022). Genomic prediction models for traits differing in heritability for soybean, rice, and maize. *BMC Plant Biology*, *22*, Article 87. https://doi.org/10.1186/s12870-022-03479-y

Karaman, E., Su, G., Croue, I., & Lund, M. S. (2021). Genomic prediction using a reference population of multiple pure breeds and admixed individuals. *Genetics, Selection, Evolution*, *53*, Article 46. https://doi.org/10.1186/s12711-021-00637-y

Kolawole, A. O., Menkir, A., Gedil, M., Blay, E., Ofori, K., & Kling, J. G (2017). Genetic divergence in two tropical maize composites after four cycles of reciprocal recurrent selection. *Plant Breeding*, *136*, 41–49. https://doi.org/10.1111/pbr.12439

Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., & Li, X. (2008). Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. *Genetics*, *180*, 1725–1742. https://doi.org/10.1534/genetics.108.091942

Li, X., Zhu, C., Wang, J., & Yu, J. (2012). Computer simulation in plant breeding. In D. L. Sparks (Ed.), *Advances in agronomy* (pp. 219–264) Academic Press.

Li, Y., Shi, F., Lin, Z., Robinson, H., Moody, D., Rattey, A., Godoy, J., Mullan, D., Keeble-Gagnere, G., Hayden, M. J., Tibbits, J. F. G., & Daetwyler, H. D. (2022). Benefit of introgression depends on level of genetic trait variation in cereal breeding programmes. *Frontiers in Plant Science*, *13*, 786452. https://doi.org/10.3389/fpls.2022.786452

Lorenz, A. J., & Smith, K. P. (2015). Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop Science*, *55*, 2657–2667. https://doi.org/10.2135/cropsci2014.12.0827

Luo, Z., Brock, J., Dyer, J. M., Kutchan, T., Schachtman, D., Augustin, M., Ge, Y., Fahlgren, N., & Abdel-Haleem, H. (2019). Genetic diversity and population structure of a *Camelina sativa* spring panel. *Frontiers in Plant Science*, *10*, 184. https://doi.org/10.3389/fpls.2019.00184

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (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

Meuwissen, T. H. E., Sonesson, A. K., Gebregiwergis, G., & Woolliams, J. A. (2020). Management of genetic diversity in the era of genomics. *Frontiers in Genetics*, *11*, 880. https://doi.org/10.3389/fgene.2020.00880

Muleta, K. T., Pressoir, G., & Morris, G. P. (2019). Optimizing genomic selection for a sorghum breeding program in Haiti: A simulation study. *G3 Genes|Genomes|Genetics*, *9*, 391–401. https://doi.org/10.1534/g3.118.200932

Neyhart, J. F., Tiede, T., Lorenz, A. J., & Smith, K. P. (2017). Evaluating methods of updating training data in long-term genomewide selection. *G3 Genes|Genomes|Genetics*, *7*, 1499–1510. https://doi.org/10.1534/g3.117.040550

Obšteter, J., Jenko, J., Hickey, J. M., & Gorjanc, G. (2019). Efficient use of genomic information for sustainable genetic improvement in small cattle populations. *Journal of Dairy Science*, *102*(11), 9971–9982. https://doi.org/10.3168/jds.2019-16853

Pocrnic, I., Obšteter, J., Gaynor, R. C., Wolc, A., & Gorjanc, G. (2023). Assessment of long-term trends in genetic mean and variance after the introduction of genomic selection in layers: A simulation study. *Frontiers in Genetics*, *14*, 1168212. https://doi.org/10.3389/fgene.2023.1168212

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodríguez, V. M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C.-C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., & Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, *192*, 715–728. https://doi.org/10.1534/genetics.112.141473

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., & Sorrells, M. E. (2015). Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. *The Plant Genome*, *8*, plantgenome2014.09.0046. https://doi.org/10.3835/plantgenome2014.09.0046

Sabadin, F., DoVale, J. C., Platten, J. D., & Fritsche-Neto, R. (2022). Optimizing self-pollinated crop breeding employing genomic selection: From schemes to updating training sets. *Frontiers in Plant Science*, *13*, 935885. https://doi.org/10.3389/fpls.2022.935885

Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., & Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Science*, *61*, 839–852. https://doi.org/10.1002/csc2.20377

Technow, F., Podlich, D., & Cooper, M. (2021). Back to the future: Implications of genetic complexity for the structure of hybrid breeding programs. *G3 Genes|Genomes|Genetics*, *11*(7), jkab153. https://doi.org/10.1093/g3journal/jkab153

Tsindi, A., Eleblu, J. S. Y., Gasura, E., Mushoriwa, H., Tongoona, P., Danquah, E. Y., Mwadzingeni, L., Zikhali, M., Ziramba, E., Mabuyaye, G., & Derera, J. (2023). Analysis of population structure and genetic diversity in a Southern African soybean collection based on single nucleotide polymorphism markers. *CABI Agriculture and Bioscience*, *4*, Article 15. https://doi.org/10.1186/s43170-023-00158-2

van den Berg, I., Boichard, D., & Lund, M. S (2016). Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. *Genetics, Selection, Evolution.*, *48*, Article 83. https://doi.org/10.1186/s12711-016-0259-0

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R. J., & Stahl, A. (2020). How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. *Frontiers in Plant Science*, *11*, 592977. https://doi.org/10.3389/fpls.2020.592977

Wientjes, Y. C. J., Bijma, P., Calus, M. P. L., Zwaan, B. J., Vitezica, Z. G., & van den Heuvel, J. (2022). The long-term effects of genomic selection: 1. Response to selection, additive genetic variance, and genetic architecture. *Genetics, Selection, Evolution*, *54*(1), Article 19. https://doi.org/10.1186/s12711-022-00709-7

Yabe, S., Yamasaki, M., Ebana, K., Hayashi, T., & Iwata, H. (2016). Island-model genomic selection for long-term genetic improvement of autogamous crops. *PLoS One*, *11*(4), e0153945. https://doi.org/10.1371/journal.pone.0153945

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fritsche-Neto, R., Yassue, R. M., da Silva, A. V., Prado, M., & DoVale, J. C. (2024). Elite germplasm introduction, training set composition, and genetic optimization algorithms effect on genomic selection-based breeding programs. *Crop Science*, *64*, 3323–3338. https://doi.org/10.1002/csc2.21384