frontiers in Genetics

ORIGINAL RESEARCH
published: 01 September 2021
doi: 10.3389/fgene.2021.637133

Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding

Éder David Borges da Silva[1], Alencar Xavier[2,3] and Marcos Ventura Faria[1]*

[1] Department of Agronomy, Universidade Estadual do Centro-Oeste, Guarapuava, Brazil, [2] Department of Biostatistics, Corteva Agriscience™, Johnston, IA, United States, [3] Department of Agronomy, Purdue University, West Lafayette, IN, United States

Genomic-assisted breeding has become an important tool in soybean breeding. However, the impact of different genomic selection (GS) approaches on short- and long-term gains is not well understood. Such gains are conditional on the breeding design and may vary with a combination of the prediction model, family size, selection strategies, and selection intensity. To address these open questions, we evaluated various scenarios through a simulated closed soybean breeding program over 200 breeding cycles. Genomic prediction was performed using genomic best linear unbiased prediction (GBLUP), Bayesian methods, and random forest, benchmarked against selection on phenotypic values, true breeding values (TBV), and random selection. Breeding strategies included selections within family (WF), across family (AF), and within pre-selected families (WPSF), with selection intensities of 2.5, 5.0, 7.5, and 10.0%. Selections were performed at the F4 generation, where individuals were phenotyped and genotyped with a 6K single nucleotide polymorphism (SNP) array. Initial genetic parameters for the simulation were estimated from the SoyNAM population. WF selections provided the most significant long-term genetic gains. GBLUP and Bayesian methods outperformed random forest and provided most of the genetic gains within the first 100 generations, being outperformed by phenotypic selection after generation 100. All methods provided similar performances under WPSF selections. A faster decay in genetic variance was observed when individuals were selected AF and WPSF, as 80% of the genetic variance was depleted within 28–58 cycles, whereas WF selections preserved the variance up to cycle 184. Surprisingly, the selection intensity had less impact on long-term gains than did the breeding strategies. The study supports that genetic gains can be optimized in the long term with specific combinations of prediction models, family size, selection strategies, and selection intensity. A combination of strategies may be necessary for balancing the short-, medium-, and long-term genetic gains in breeding programs while preserving the genetic variance.

Keywords: long-term gains, soybean breeding, genomic selections, selection intensity, genomic prediction

OPEN ACCESS

Edited by:
Waseem Hussain, International Rice Research Institute (IRRI), Philippines

Reviewed by:
Yongkang Kim, University of Colorado Boulder, United States
Nicholas B. Larson, Mayo Clinic, United States

****Correspondence:***
Éder David Borges da Silva
ederdbs@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 December 2020*
Accepted: *05 August 2021*
Published: *01 September 2021*

Citation:
Silva ÉDB da, Xavier A and Faria MV (2021) Impact of Genomic Prediction Model, Selection Intensity, and Breeding Strategy on the Long-Term Genetic Gain and Genetic Erosion in Soybean Breeding. Front. Genet. 12:637133. doi: 10.3389/fgene.2021.637133

INTRODUCTION

Soybean [*Glycine max* (L.)] is the most important source of protein for animal feed and an important source of oil for human consumption, biofuel, and other industrial applications. Soybeans are cultivated globally, and the largest producers include Brazil, United States, Argentina, Paraguay, and China (FAO, 2021). Soybeans are bred for several traits, but grain yield is considered as the most important.

Genome-wide prediction is a key tool in soybean breeding. It is utilized for faster and more accurate selection of superior individuals (Meuwissen et al., 2001). Methodologically, genomic models recreate the framework utilized for pedigree analysis, but using genomic relationships instead (VanRaden, 2008; Habier et al., 2011; VanRaden et al., 2011). Other factors that may have contributed to the increasing adoption of genomic selection (GS) in plants include the decreasing cost of genotyping and the availability of software tools and computing power to analyze large datasets.

Studies involving GS in plants have been mostly focused on prediction for advancement purposes, hence restricted to the evaluation of genetic gain within a single generation (Schmutz et al., 2010; Sonah et al., 2013; Jarquin et al., 2016; Xavier et al., 2016, 2018a,b; Diers et al., 2018; Smallwood et al., 2019). Studies of long-term gains based on GS are expensive and time-consuming; consequently, the literature is scarce (Wray and Goddard, 1994; Goddard, 2009; Yabe et al., 2016; Gorjanc et al., 2018; Allier et al., 2019a). In addition, evaluation with real data from breeding programs faces additional challenges, such as the ongoing changes in breeding pipelines driven by business decisions, changes in the genotyping technology, and annual changes in resources. Conversely, the deployment of simulations has become an instrumental decision tool in plant breeding. It enables the assessment of genetic gain under different scenarios. In part, the increasing popularity of simulations is due to the quantity and flexibility of software made available (Faux et al., 2016; Pook et al., 2019; Toledo et al., 2019). For instance, breeders are now capable of simulating entire breeding programs with the intent of tuning the breeding parameters to maximize genetic gains in the short and long term (Hickey et al., 2014; Gorjanc et al., 2018), along with the best allocation of resources for a given budget.

By assessing predictive models and contrasting selection strategies, this study envisioned analyzing the influence of a set of variables on long-term genetic gains based on a simulated soybean breeding program and providing insight into the best practices for optimizing genetic gains.

MATERIALS AND METHODS

Simulated Populational Parameters

The founder breeding population contained 200 individuals. Those were simulated based on the genomic parameters using the Markovian Coalescent Simulator (MaCS; Chen et al., 2009), which recreates the evolutionary process with multiple cycles of drift, mutation, and selection. The genomic parameters for the simulations reproduce the soybean genome with detailed information (Schmutz et al., 2010). We considered a genetic map architecture of 20 chromosomes with 115 cM average length, which collectively spanned 950 Mb. For each chromosome, 1,000 segregating sites were assigned.

Our study focused on the simulation of grain yield (in tons per hectare) as the primary trait of interest. The genetic architecture of the simulated trait was assumed to be infinitesimal with 70% of all segregating sites, which were not necessarily utilized as markers, having a non-zero effect sampled from a normal distribution. The genotype-by-environment variance provided a non-heritable variation attributed to the season. Residual variance remained constant throughout the simulation, causing a reduction in heritability overtime as the genetic variance decreased. Simulations began assuming an average yield of 3.00 t ha^{-1}. The function *addTraitAEG* from the AlphaSimR package was utilized for the simulation of the phenotypic values. All simulation code is available on GitHub.[1]

Additive genetic effects, genotype-by-environment interaction, and residuals were simulated from Gaussian distribution using variance components estimated from the SoyNAM dataset (Diers et al., 2018; Xavier et al., 2018a) as $\sigma^2_a = 25$, $\sigma^2_{G\times E} = 49$, $\sigma^2_e = 121$, and $h^2 = 0.12$. The parameter estimation from the SoyNAM dataset was based on a multivariate genomic best linear unbiased prediction (GBLUP) model with unstructured genetic covariance and diagonal residual covariance, fitting grain yield from all 18 environments as response variables and using as explanatory variables the overall mean (fixed) and a polygenic term (random). The final estimates of the variance components for σ^2_a, σ^2_e, and h^2 were obtained as averages across the 18 environments, whereas $\sigma^2_{G\times E}$ was computed as the average off-diagonal of the variance–covariance matrix.

The main simulation settings followed a soybean breeding program with 300 families per cycle and with 50 individuals per family, producing a total of 15,000 individuals per cycle. After crossing, the populations were inbred *via* single seed descendent (SSD) until $F_{2:4}$, as shown in **Figure 1**, where lines were evaluated in field trials and genotyped with a single nucleotide polymorphism (SNP) array similar to the Soybean 6K SNP chip (Akond et al., 2013). Individuals were then selected to become parents of the upcoming breeding cycle using the phenotypic and genotypic information. The calibration of genomic prediction leveraged data from the previous three breeding cycles, thus leveraging information from up to 45,000 individuals per model. The processes of selecting and crossing were repeated for 200 cycles to capture the theoretical plateau of genetic gains across all simulated parameters. Each breeding scenario was reproduced 60 times with different computational random seeds.

A second simulation with 100 breeding cycles was performed with varying numbers of families and offspring, where five combinations that use the same number of resources were chosen—300 × 50, 250 × 60, 200 × 75, 150 × 100, and 100 × 150—where the combinations correspond to the number

[1]https://github.com/Ederdbs/GenomicSelection

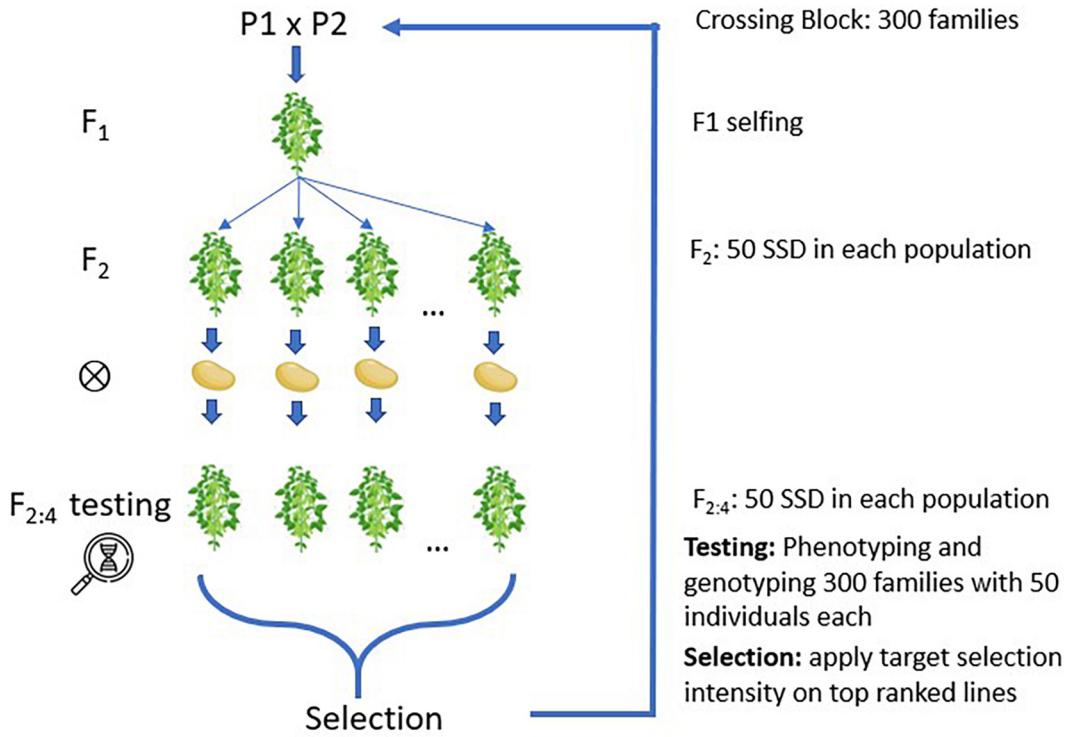

FIGURE 1 | Simulated families created and inbreeding using single seed descent (SSD).

of families and individuals per family, respectively. Each breeding scenario was reproduced 45 times with different random seeds.

Genotypic and phenotypic data were simulated with the R package AlphaSimR (Gaynor et al., 2020), reproducing the previous methodological framework (Faux et al., 2016). The software was utilized to simulate the founder population, perform selection, fingerprint individuals with the specified SNP chip, make crosses, generate offspring, inbred individuals, and simulate phenotypic values. All simulations and subsequent statistical analyses of the results were performed using R software (R Core Team, 2020). The code was run in parallel by distributing the multiple breeding scenarios over 960 cores, requiring approximately 10 h of computation per run. The R package doParallel (Ooi et al., 2019) was utilized to parallelize the runs.

Evaluation of Simulated Scenarios

Evaluation of the breeding strategies, selection intensities, and selection models was based on previous studies (Daetwyler et al., 2013). The evaluation criteria included the population mean across breeding cycles, genetic variance, and accuracy. Analyses were performed within a generation, combining the data from the repeated simulation runs. The statistical model for the analysis of simulated data was the following:

$$y = 1\mu + \mathbf{X}_m m + \mathbf{X}_s s + \mathbf{X}_i i + \mathbf{X}_p p + \varepsilon$$

where y is the vector of the random variable of the simulated population; μ is the model intercept; $\mathbf{X}$ represents the incidence matrix, which is further divided to accommodate the three factors under evaluation ($\mathbf{X}_m$, $\mathbf{X}_s$, $\mathbf{X}_i$, and $\mathbf{X}_p$); m for the selection model; s for the breeding strategy; i for the selection intensity; p for the population design, as combinations of the number of families and individuals per family; and ε is the vector of residuals, assumed to be distributed as $\varepsilon \sim N\left(0, I\sigma_\varepsilon^2\right)$. The statistical test of multiple comparison was based on Tukey's range test with 5% probability of error fit using the built-in R function TukeyHSD. This model was used to generate **Figure 2**.

Selection Models

The following selection models are evaluated: (1) True breeding values (*TBV*)—true breeding value, which serves as the upper limit of the achievable prediction power; (2) *Random*—random selection of individual, as the worst-case scenario; (3) *Pheno*—phenotypic-based selection without the use of genomic information; (4) *GBLUP*—the genomic best linear unbiased predictor fitted with REML (restricted maximum likelihood) variance components (Nejati-Javaremi et al., 1997; Habier et al., 2007); (5) *BayesA*—Bayesian shrinkage regression that assigns a t prior to marker effects (Meuwissen et al., 2001); (6) *BayesB*—an extension of BayesA with variable selection (Meuwissen et al., 2001); (7) *FLM*—fast Laplace model (Xavier, 2019), an empirical Bayes model with a double exponential prior for marker effects; and (8) *RF*—random forest regression (Breiman, 2001), a common machine learning procedure based on bootstrapping aggregation of multiple decision trees. The models GBLUP, BayesA, BayesB, and FLM were fitted using

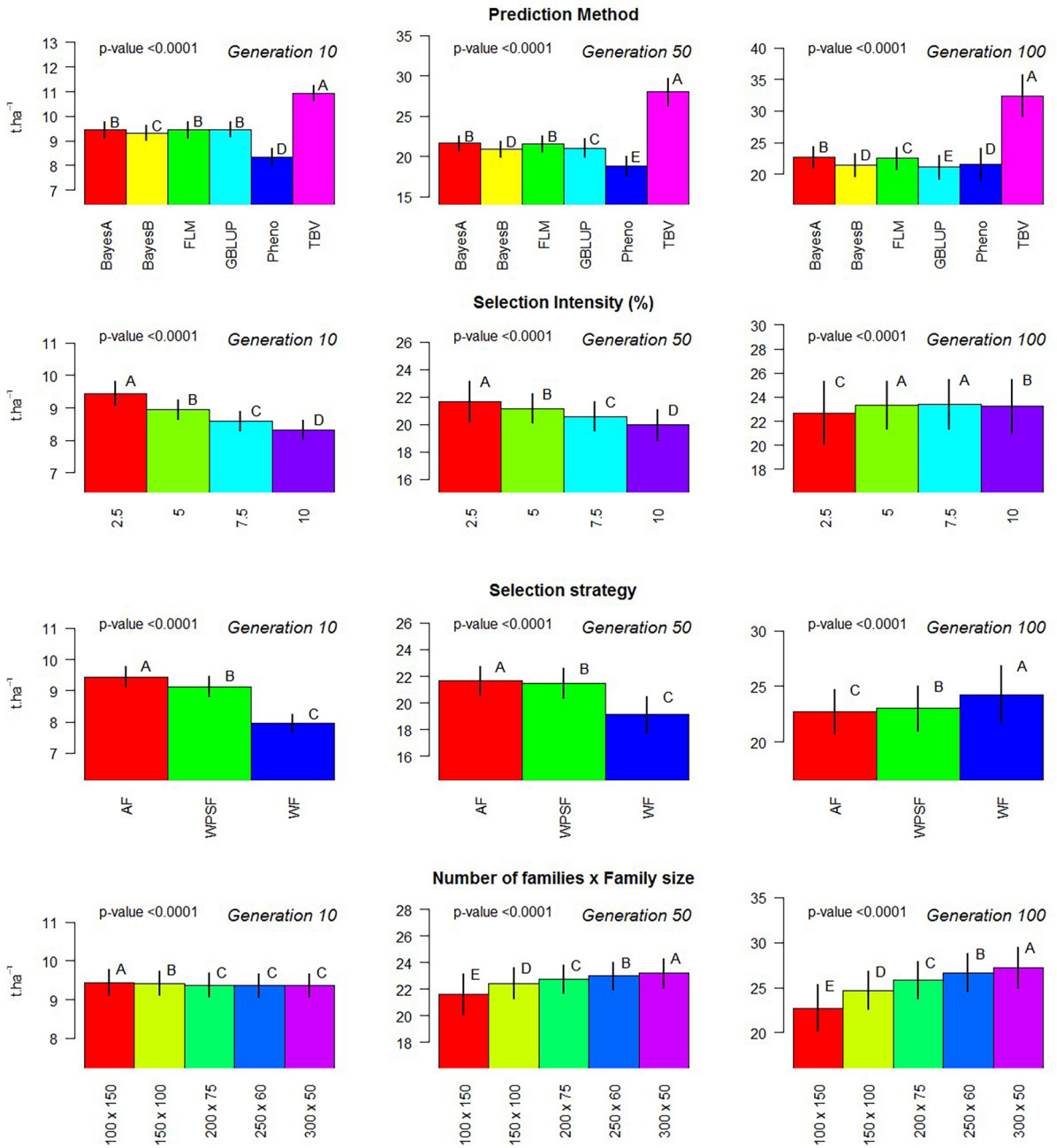

FIGURE 2 | Evaluation of individual factors on population means. Multiple comparison test: Capital letters indicate difference in means across factor with Tukey's range test with 5% alpha level contrasting the levels of each factor (prediction method, selection intensity, balance between population size and family size, and breeding strategy) on generations 10, 50, and 100.

the R package bWGR (Xavier et al., 2019) and solved *via* expectation–maximization (EM). The model RF was fitted using the R package ranger (Wright et al., 2020) with default settings.

As a brief description of the GS model, these models in function on genomic information can be written in terms of the linear model:

$$y = \mathbf{X}b + f(\mathbf{M}) + \varepsilon$$

where y is the vector of phenotypic values; $\mathbf{X}$ is the incidence matrix of the environment term treated as a fixed effect; b is a vector of environmental means; $f(\mathbf{M})$ is the function of markers that describe the genetic merit of individuals; and ε is a random vector of residuals, assumed to be distributed as $\varepsilon \sim N\left(0, \mathbf{I}\sigma_{\varepsilon}^{2}\right)$. The genetic function of markers, $f(\mathbf{M})$, varied from model to model. For GBLUP, BayesA, and FLM, the function was linear and the marker effects were strictly additive; thus, the function of markers was $f(\mathbf{M}) = \mathbf{M}\beta$. The distinction of the models was the

prior assigned to the distribution of marker effects, being normal for GBLUP, distributed as Student's t for BayesA, and distributed as a double exponential for FLM. The function describing BayesB was $f(\mathbf{M}) = \mathbf{M}\beta\gamma$, which is also linear, but with a variable selection term (γ) that caused further shrinkage to the Student's t prior assigned to the marker effects. The only non-linear model under evaluation was random forest, in which case the genetic function is a linear ensemble of multiple independent regression trees (T): $f(\mathbf{M}) = n^{-1} \sum T(m \in \mathbf{M})$.

Breeding Strategy

The breeding strategies were based on soybean breeding designs previously described in the literature (Backes et al., 2003; Sebastian et al., 2010; de Cássia Pereira et al., 2017; da Silva et al., 2018; Smallwood et al., 2019). The following approaches were considered in this study:

AF: across-family selection. Genotypes are selected across families based on their estimated genetic merit, without regard for their family structure or any constraint for selecting multiple individuals from the same pedigree.

WF: within-family selection. In this strategy, all families were equally represented in the advancements. The best genotypes from each family are selected to become parents in the upcoming generations.

WPSF: within the pre-selected family. This strategy comprises two steps. Firstly, the family level selection is performed to identify the best-performing families (top 30%). Secondly, the selection of individuals occurs within the family. With fewer families to select from, more individuals per family will be parenting the upcoming generation compared to WF.

Selection Intensity

Four levels of selection intensity were considered: 2.5, 5.0, 7.5, and 10.0%. These values represent the percentages of individuals selected to be used as parents of the next generation. The selection of parental combinations was performed at random; thus, it is possible that not all selected individuals served as parents.

RESULTS

Genetic Gains

The simulation results presented in **Figure 3** summarize the population means over the course of 200 cycles. **Supplementary Table 1** provides the population means for all combinations of treatments under evaluation in breeding cycles 10, 100, and 200. Across all scenarios, the population mean of random selection is anchored at the starting point. Selection of TBV represents the upper boundary of each scenario; hence, these are particularly useful to contrast the potential of the different scenarios. The highest long-term population means from selection on TBV occurred WF with loose selection intensities (7.5–10%). Genetic gains were generally closer to those from TBV when selections were performed WPSF.

Phenotypic selection outperformed GS over the course of 200 breeding cycles. Selection using random forest provided poor predictive performance in all scenarios, possibly due to the non-additive nature of the regression trees fitting a strictly additive genetic architecture. All linear genomic models (BayesA, BayesB, FLM, and GBLUP) provided similar outcomes. When conditioning for all other varying parameters, BayesA and FLM were the best-performing models within the first 100 breeding cycles (**Figure 2**).

After 10 cycles of selection, the highest gains were attained at the highest selection intensity (2.5%), which characterizes the short-term gain benefit from a higher selection pressure while the genetic variance is still abundant. After 100 breeding cycles, the genetic gains are affected by the combination of selection intensity and breeding strategy. For example, selection performed AF using BayesA provided the highest gains with a selection intensity of 10%, whereas, under WF, the highest gains occurred with a selection intensity of 2.5%. Such discrepancy is attributed to the amount of genetic variance left for long-term selection.

The highest long-term gains were reached when selections were performed WF. The maximum attainable, as benchmarked by selection upon TBV, resulted in a grain yield of 54 t ha^{-1} (WF), being 35% higher than AF selections and 46% higher than WPSF (**Supplementary Table 2**). The overall trend for long-term gains using GS followed the order WF > WPSF > AF. When the selections were based on phenotypic values, the genetic gains outpaced the GS run for all strategies (AF, WPSF, and WF), whereas that was not observed within the first 100 cycles (**Figure 1**). In fact, phenotypic selection WF was the third highest performing model, behind AF and WF selections performed on TBVs. The impact of each factor on the prediction accuracy over 200 breeding cycles is provided in **Supplementary Figure 1**.

Figure 2 summarizes the results of the simulation performed within 100 cycles, where different family sizes were an additional variable under evaluation. Within 10 breeding cycles, the scenario of 100 families with 150 individuals displayed the highest average, although the differences were negligible. Over the course of 50 and 100 breeding cycles, the number of families and the family sizes displayed significant differences in the genetic gains, with larger differences as generations progressed. The overall trend was that a greater number of families increase the gain in the long term.

Diversity Loss

The decay in genetic variance overtime is presented in **Figure 4**. The number of cycles to exhaust 80% of the genetic variance is provided in **Supplementary Table 3**. The study simulates closed populations without the inflow of external variation, the existing genetic variance consumed overtime as selection takes place. Overall, a fast decay in genetic variance is observed under a higher selection pressure, whereas a lower selection pressure preserved more genetic variance in the long term. When selection was performed at random, over 80% of the initial genetic variance remained after 200 breeding cycles. The interaction between the selection intensity and selection strategy was significant ($p < 0.01$) across all selection models.

Within-family selection preserved the genetic variance for more cycles (**Figure 4**). Selection WF based on TBVs exhausted 80% of the genetic variance within 48–69 breeding cycles, whereas AF and WPSF selections on TBVs exhausted 80% of

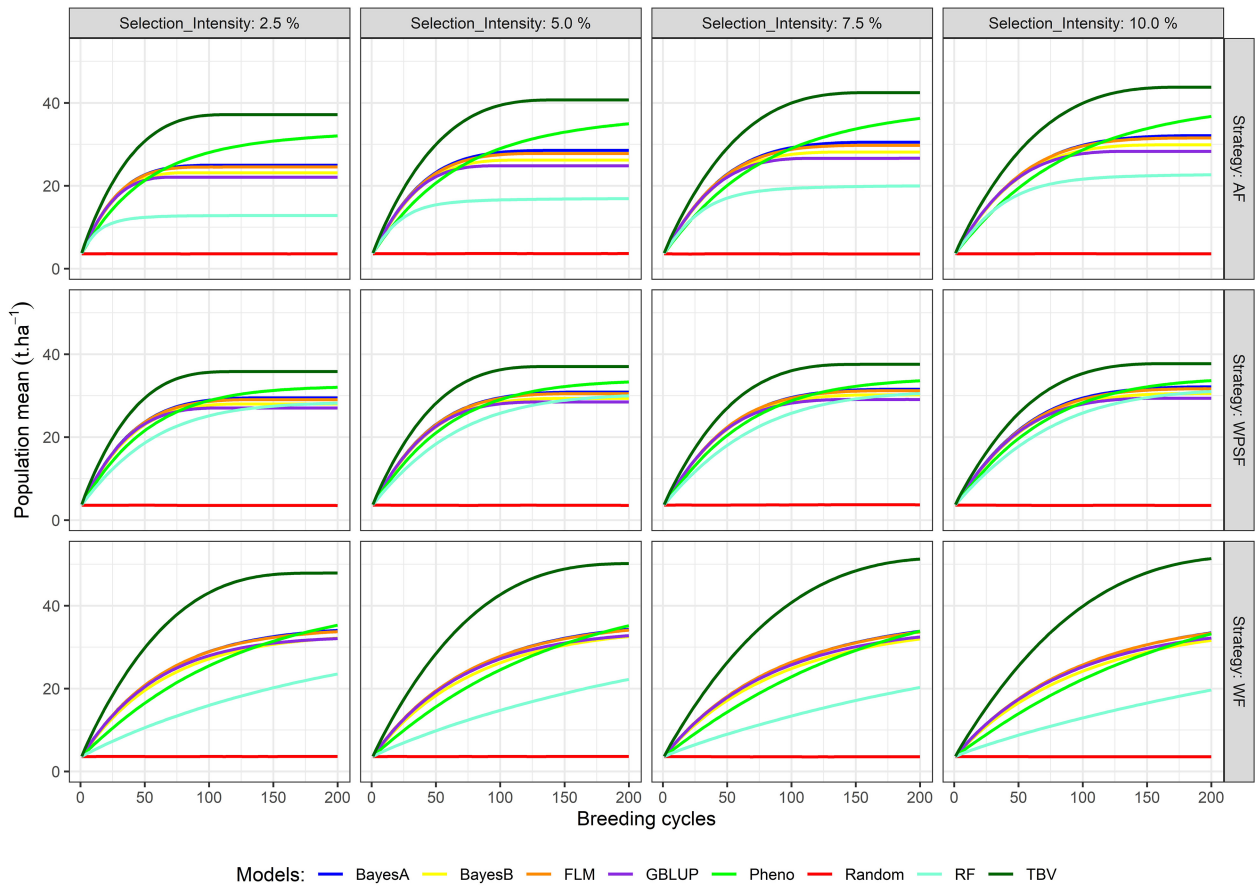

FIGURE 3 | Population means across 200 breeding cycles. Colors correspond to the different selection methods, grid columns represent the selection intensity (2.5, 5, 7.5, and 10%), and grid rows represent the breeding strategies, where individuals were selected across family (AF), within pre-selected families (WPSF), or within-family (WF).

the diversity between 25 and 42 cycles (**Supplementary Table 3**). Depletion of genetic variance was more pronounced with GS. Under the selection intensity of 10%, BayesA selection WF exhausted 80% of the variance after 184 cycles, whereas selections AF and WPSF display the same diversity loss after 54 and 58 cycles, respectively.

Diversity loss attributed to genetic drift is presented in **Figure 5**. These results assess the impact of bottlenecking the population through the various combinations of breeding strategy and selection intensity, utilizing random selections to avoid the confounding effect of directional selection. Higher rates of drift occurred under a higher selection pressure (2.5%). Strategy-wise, losses were highest for selection WPSF, with little difference across the selection intensities, ranging from −0.325 to −0.353%. The lowest rate of drift was observed under WF selection, with the rate of losses ranging from −0.199 to −0.136%.

DISCUSSION

Genomic prediction has become an important tool for selection and breeding in agriculture as it can enhance the rate of genetic gain in comparison to pedigree and phenotype-based selection by leveraging information on relationship and the linkage disequilibrium between the marker and the quantitative trait locus (QTL; Meuwissen et al., 2001; Habier and Fernando, 2009; Bernardo, 2010; Crossa et al., 2013, 2017; Daetwyler et al., 2013; de Los Campos et al., 2013). In soybean, the value of genomic prediction has been assessed and described in recent years (Jarquín et al., 2014; Xavier et al., 2016, 2018a,b; Diers et al., 2018; Matei et al., 2018; Xavier and Rainey, 2020). These studies agreed that adequate composition of the training data is imperative to successful and accurate prediction. The definition of an optimized training set entails (1) maximizing the genetic relationship between the training and target populations and (2) collecting phenotypic information from year–location combinations that represent the target population of environments. Whereas factors that affect genomic predictions for short-term gains have been well characterized, it is unclear which factors affect long-term genetic gains. The answer for that would come from long-term simulations, such as the present study. Primarily, simulations enable the optimization of the modern breeding program in animal and plant species (Yu et al., 2005; Hickey et al., 2014; Cowling et al., 2015, 2020;

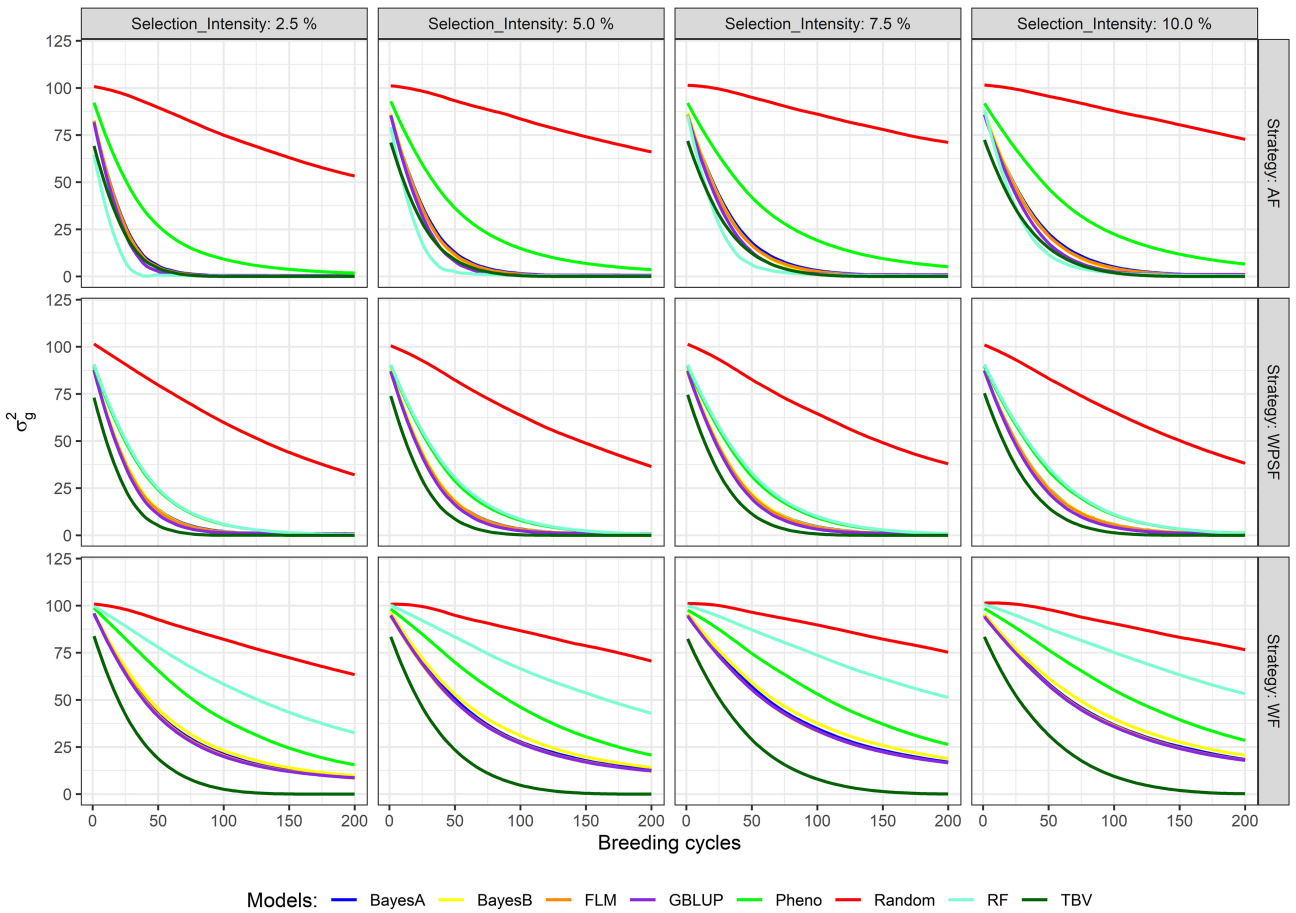

FIGURE 4 | Genetic variance across 200 breeding cycles. Colors correspond to the different selection methods, grid columns represent the selection intensities (2.5, 5, 7.5, and 10%), and grid rows represent the breeding strategies, where individuals are selected across family (AF), within pre-selected families (WPSF), or within family (WF).

Gorjanc and Hickey, 2018; Muleta et al., 2018) by enabling the assessment of the breeding conditions that increase the rate of genetic gains, the conservation of useful genetic diversity, and the best allocation of breeding resource, such as the number of field plots, genotyping density, number of crosses, and population size (Heffner et al., 2010; Gonen et al., 2017; Gorjanc et al., 2017a,b).

Simulations indicate that linear models outperformed random forest for complex traits controlled by additive genetics and additive genotype-by-environment interactions. Under different scenarios, other studies found machine learning methods to display similar performances (Li et al., 2018; Ali et al., 2020). The discrepancy in the results is likely due to the nature of trait and population under evaluation, as machine learning predictions could be suitable for more structured populations and with some degree of epistatic control (Xavier, 2019; Abdollahi-Arpanahi et al., 2020). We also acknowledge that random forest was run with default settings in this study, and parameter tuning would benefit its predictive performance.

Selection factors provided a similar outcome to the findings in other studies (Gorjanc and Hickey, 2018; Santantonio and Robbins, 2020), where the authors assessed balancing short- and long-term sustainable gains in plant breeding. Their results indicate that higher population sizes provide higher long-term gains. An alternative framework for the maximization of long-term response to selection is proposed by Goddard (2009) based on the use of selection indexes that account for allele frequency aiming to account for the value of rare loci and in short- and long-term gains. Under limited resources, our simulations indicate that a lower selection pressure generally contributes to long-term gains at the cost of compromising short-term gains. Across breeding strategies, WPSF appears to provide reasonable gains in both the short and the long term while having the range of gains being less influenced by selection pressure. WPSF is an intermediate between AF and WF, and the results are, in fact, intermediary between the short-term gains provided by AF selections and the long-term gains provided by WF selection.

The real-life trend of genetic gains in soybeans is positive, but variable across geographies. In North America, the rates of genetic gain have been estimated to be 23.4 kg ha^{-1} $year^{-1}$ (Fox et al., 2013), 26.5 kg ha^{-1} $year^{-1}$ (Koester et al., 2014), and 16.8 kg ha^{-1} $year^{-1}$ (Rogers et al., 2015). In the southern regions of Brazil, the rates of genetic gains were estimated to be

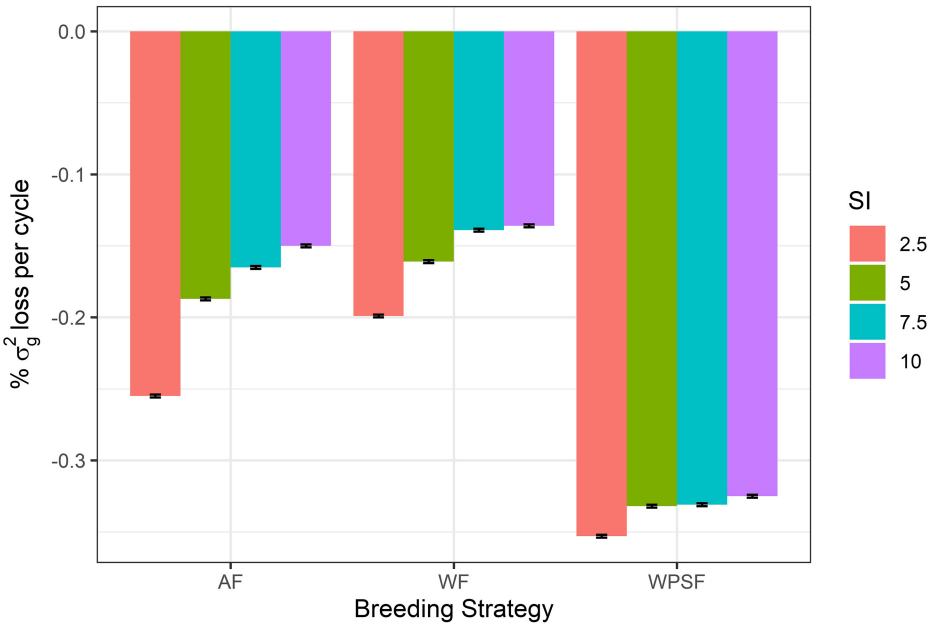

FIGURE 5 | Genetic drift per cycle under random selection across family (AF), within pre-selected families (WPSF), or within family (WF) in different selection intensities (SI).

71.5 kg ha^{-1} year^{-1} (Lange and Federizzi, 2009) and 40.0 kg ha^{-1} year^{-1} (Todeschini et al., 2019); in Argentina, the rate has been reported to be 44.3 kg ha^{-1} year^{-1} (de Felipe et al., 2016). These reports provide insight from the perspective of traditional breeding progress before the deployment of GS and, in most cases, with lengthy breeding cycles with the choice of parents taking place in advanced generations and commercial products. Our simulations provided higher annual gains than what has been reported; however, with the advent of earlier evaluations and increasing trust in genomic prediction, it is likely that annual genetic gains will be progressively and iteratively optimized for multiple factors, including those evaluated in the present study (model, selection intensity, family size, and breeding strategy).

The selection of unproven parents from earlier generations is often interpreted as gambling with high risk and high rewards, even though much of the risk is mitigated with the use of genomic information with robust statistical models calibrated with phenotypic data from multiple years. In addition to advancements, more opportunities arise with the use of genomics to predict and select the best combinations for crossing that further increase the probability of generating elite offspring. Previous studies have evaluated population-level selection strategies in further detail (Bernardo, 2010; Jannink, 2010; Kemper et al., 2012; Daetwyler et al., 2015; Ma et al., 2016; Goiffon et al., 2017; Matei et al., 2018) with the goal of preserving the segregation of low-frequency haplotypes for long-term gains (Beukelaer et al., 2017). Balancing the number of families and the family size can be a fundamental part of the strategy to continue the steady gains overtime (**Figure 2**), and, whereas the difference is not perceived in the short term, the magnitude of grain increases significantly overtime. Yet, multiple factors should be taken into account when allocating resources in terms of the number of families and family size (Lindgren et al., 1997; Fu, 2015).

Scenarios simulated as provided herein were based on the parental selection at the F4 stage, which is commonly perceived as an early generation for recycling as the quality and the quantity of phenotypic data are still scarce, of doubtful quality, and in many cases, without replication. Nevertheless, early recycling is a promising framework for speeding up the rate of genetic gain by shortening the length of the breeding cycles. In fact, shortening the breeding cycles while inducing multiple cycles a year reproduces a framework referred to as "speed breeding" (Hickey et al., 2019; Nagatoshi and Fujita, 2019; Jähne et al., 2020). Recent studies often support recombination in the early stages of inbred development (Gaynor et al., 2017), more so as the accuracy of selection in the early stages benefits greatly from the GS. Another important aspect of parental selection regards the management of genetic diversity in modern plant breeding, which is largely ignored and not always adequately measured (Fu, 2015). Our results indicate that the multiple factors in the breeding design can affect the rate of diversity loss, mainly selection pressure and selection strategy (**Supplementary Table 2**), and that one must consider to balance these factors to attain the desired gain in the short term without compromising long-term gains. That is particularly the case for soybeans, whose germplasm-wise genetic diversity is considered low when compared to that of other species (Martin, 1982). Some Canadian soybean breeding programs have maintained diversity through decades of breeding while fixing maturity genes (Bruce et al., 2019). In the United States, soybean population structures and diversity varied by maturity group (Vaughn and Li, 2016), which

suggests that new sources of variation could be obtained through the introgression of material from different regions.

The diversity available in breeding programs affects the accuracy of breeding values by dictating the amount of existing genetic signals to select upon an effective population size (Meuwissen, 2009). With restricted diversity, the genotyping density and marker distribution can be optimized to capture the existing variation in the target population with the goal of increasing genomic prediction accuracy (Ma et al., 2016). Of course, the long-term impacts of selection on genetic variance also vary depending on the genetic variance of interest, as the prominence of additive and non-additive variances is not the same over multiple cycles of selection (Paixão and Barton, 2016).

In soybeans, the management of diversity is necessary to ensure useful variability for future breeding objectives, such as yield performance under drought or waterlogging (Valliyodan et al., 2017), the seed oil and protein content profiles (Stewart-Brown et al., 2019), and disease resistance (de Azevedo Peixoto et al., 2017). Monitoring genetic diversity in the genomic era can be performed through tracking overtime changes in allele frequencies (Allier et al., 2019b; de Castro Lara et al., 2020; Meuwissen et al., 2020). We showed that selection could quickly exhaust genetic diversity under closed breeding systems, and breeding systems can benefit from balancing short gains to preserve diversity and assure long-term gains. Such balance had been the focal point of recent studies (Cowling et al., 2017; Gorjanc et al., 2018; Ru and Bernardo, 2019, 2020; Santantonio and Robbins, 2020) seeking for avenues to extend genetic resources with genomic tools, including the selection of material from germplasm collection to expend the genetic basis of elite programs. In addition to germplasm introgression, increases in genetic diversity in soybeans have been done in the past through mutagenic agents (Curtin et al., 2011; Khan, 2013; Haun et al., 2014; Demorest et al., 2016) and more recently, through genome editing techniques based on CRISPR-Cas9 (Cai et al., 2015, 2018a,b; Jacobs et al., 2015; Sun et al., 2015; Zheng et al., 2020) and target recombination for directional backcrossing (Ru and Bernardo, 2019, 2020).

The simulations performed in our study indicate that GS enables higher rates of genetic gain in the short and medium term compared with phenotype selection, but also led to faster extinction of the genetic variance. Thus, genomic prediction and selection must be applied mindfully with the purpose of maximizing gains while maintaining genetic variance. We found that a breeding strategy that balances selection at the family level, and within and across family at the individual level, can mitigate losses in genetic variance while providing satisfying genetic gains in the short term. Simulation is a powerful and inexpensive tool to test hypotheses, and for future studies, we envision addressing the importance of other important breeding parameters. Namely, future studies should focus on investigating (1) the optimal generation to select the parents and its trade-off with the accuracy of selection; (2) the influence of non-additive and non-infinitesimal genetic architecture and how machine learning would perform in such conditions; (3) the long-term effect of different models designed to select parental combinations; (4) the impact of different island models where new sources of variation are constantly infused into the main breeding panel; and (5) what would be the potential benefit of breeding hybrid soybeans assuming there are variable levels of dominance.

CONCLUSION

Long-term gains were influenced by the interaction among GS models, breeding strategy, and selection intensity. Adequate handling of these factors will aid breeding programs to ensure genetic gains in short, medium, and long term. Therefore, the breeding strategy is the most influential factor and, therefore, is a key criterion to conserve genetic variance and obtain the highest population mean overtime. The absolute impact of the selection intensity is lower than that of the breeding strategy and GS model. The benefits of balancing family size and the number of families were not perceived on short-term gains. Additive GS models (BayesA, BayesB, FLM, and GBLUP) have similar behaviors in selecting the best individuals, whereas RF has poor predictive performance when implemented with default settings. In summary, a combination of strategies may be necessary for balancing the short-, medium-, and long-term genetic gains in breeding programs while preserving genetic variance.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/Ederdbs/GenomicSelection.

AUTHOR CONTRIBUTIONS

ÉS and AX implemented the research, contributed with ideas to the algorithms, and wrote the manuscript. MF implemented the research, contributed ideas, and wrote the manuscript. All authors approved the final version to be published and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

The authors are grateful to the Universidade Estadual do Centro-Oeste (UNICENTRO), because this manuscript is part of the thesis of the ÉS and to Corteva Agriscience for support in computational resources. The authors also thank the reviewers for their helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.637133/full#supplementary-material

REFERENCES

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evolut.* 52:12. doi: 10.1186/s12711-020-00531-z

Akond, M., Liu, S., Schoener, L., Anderson, J. A., Kantartzi, S. K., Meksem, K., et al. (2013). A SNP-Based genetic linkage map of soybean using the SoySNP6K Illumina infinium BeadChip genotyping array. *Plant Genet. Genomics Biotechnol.* 1, 80–89. doi: 10.5147/pggb.v1i3.154

Ali, M., Zhang, L., DeLacy, I., Arief, V., Dieters, M., Pfeiffer, W. H., et al. (2020). Modeling and simulation of recurrent phenotypic and genomic selections in plant breeding under the presence of epistasis. *Crop J.* 8, 866–877. doi: 10.1016/j.cj.2020.04.002

Allier, A., Lehermeier, C., Charcosset, A., Moreau, L., and Teyssèdre, S. (2019a). Improving short- and long-term genetic gain by accounting for within-family variance in optimal cross-selection. *Front. Genet.* 10:1006. doi: 10.3389/fgene.2019.01006

Allier, A., Teyssèdre, S., Lehermeier, C., Claustres, B., Maltese, S., Melkior, S., et al. (2019b). Assessment of breeding programs sustainability: application of phenotypic and genomic indicators to a North European grain maize program. *Theor. Appl. Genet.* 132, 1321–1334. doi: 10.1007/s00122-019-03280-w

Backes, R. L., Reis, M. S., Cruz, C. D., Sediyama, T., and Sediyama, C. S. (2003). Correlation estimates and assessment of selection strategies in five soybean populations. *CBAB* 3, 107–116. doi: 10.12702/1984-7033.v03n02a03

Bernardo, R. (2010). Genomewide selection with minimal crossing in self-pollinated crops. *Crop Sci.* 50, 624–627. doi: 10.2135/cropsci2009.05.0250

Beukelaer, H. D., Badke, Y., Fack, V., and Meyer, G. D. (2017). Moving beyond managing realized genomic relationship in long-term genomic selection. *Genetics* 206, 1127–1138. doi: 10.1534/genetics.116.194449

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F., Eskandari, M., and Rajcan, I. (2019). Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. *Theor. Appl. Genet.* 132, 3089–3100. doi: 10.1007/s00122-019-03408-y

Cai, Y., Chen, L., Liu, X., Guo, C., Sun, S., Wu, C., et al. (2018a). CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean. *Plant Biotechnol. J.* 16, 176–185. doi: 10.1111/pbi.12758

Cai, Y., Chen, L., Sun, S., Wu, C., Yao, W., Jiang, B., et al. (2018b). CRISPR/Cas9-mediated Deletion of large genomic fragments in soybean. *Int. J. Mol. Sci.* 19:3835. doi: 10.3390/ijms19123835

Cai, Y., Chen, L., Liu, X., Sun, S., Wu, C., Jiang, B., et al. (2015). CRISPR/Cas9-mediated genome editing in soybean hairy roots. *PLoS One* 10:e0136064. doi: 10.1371/journal.pone.0136064

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

Cowling, W. A., Gaynor, R. C., Antolín, R., Gorjanc, G., Edwards, S. M., Powell, O., et al. (2020). In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. *Sci. Rep.* 10:4037. doi: 10.1038/s41598-020-61031-0

Cowling, W. A., Li, L., Siddique, K. H. M., Henryon, M., Berg, P., Banks, R. G., et al. (2017). Evolving gene banks: improving diverse populations of crop and exotic germplasm with optimal contribution selection. *J. Exp. Bot.* 68, 1927–1939. doi: 10.1093/jxb/erw406

Cowling, W. A., Stefanova, K. T., Beeck, C. P., Nelson, M. N., Hargreaves, B. L. W., Sass, O., et al. (2015). Using the animal model to accelerate response to selection in a self-pollinating crop. *G3* 5, 1419–1428. doi: 10.1534/g3.115.018838

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 3, 1903–1926. doi: 10.1534/g3.113.008227

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Curtin, S. J., Zhang, F., Sander, J. D., Haun, W. J., Starker, C., Baltes, N. J., et al. (2011). Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiol.* 156, 466–473. doi: 10.1104/pp.111.172981

da Silva, F. M., De MatosPereira, E., Val, B. H. P., Perecin, D., Mauro, A. O. D., Unêda-Trevisoli, S. H., et al. (2018). Strategies to select soybean segregating populations with the goal of improving agronomic traits. *Acta Scientiarum. Agronomy* 40:39324. doi: 10.4025/actasciagron.v40i1.39324

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341–1348. doi: 10.1534/genetics.115.178038

de Azevedo Peixoto, L., Moellers, T. C., Zhang, J., Lorenz, A. J., Bhering, L. L., Beavis, W. D., et al. (2017). Leveraging genomic prediction to scan germplasm collection for crop improvement. *PLoS One* 12:e0179191. doi: 10.1371/journal.pone.0179191

de Cássia Pereira, F., Bruzi, A. T., de Matos, J. W., Rezende, B. A., Prado, L. C., and Nunes, J. A. R. (2017). Implications of the population effect in the selection of soybean progeny. *Plant Breed.* 136, 679–687. doi: 10.1111/pbr.12512

de Castro Lara, L. A., Pocrnic, I., Gaynor, R. C., and Gorjanc, G. (2020). Temporal and genomic analysis of additive genetic variance in breeding programmes. *bioRxiv* [Preprint]. doi: 10.1101/2020.08.29.273250

de Felipe, M., Gerde, J. A., and Rotundo, J. L. (2016). Soybean Genetic gain in maturity Groups III to V in argentina from 1980 to 2015. *Crop Sci.* 56, 3066–3077. doi: 10.2135/cropsci2016.04.0214

de Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193, 327–345. doi: 10.1534/genetics.112.143313

Demorest, Z. L., Coffman, A., Baltes, N. J., Stoddard, T. J., Clasen, B. M., Luo, S., et al. (2016). Direct stacking of sequence-specific nuclease-induced mutations to produce high oleic and low linolenic soybean oil. *BMC Plant Biol.* 16:225. doi: 10.1186/s12870-016-0906-1

Diers, B. W., Specht, J., Rainey, K. M., Cregan, P., Song, Q., Ramasubramanian, V., et al. (2018). Genetic architecture of soybean yield and agronomic traits. *G3* 8, 3367–3375. doi: 10.1534/g3.118.200332

FAO (2021). *FAO Global Statistical Yearbook, FAO Regional Statistical Yearbooks*. Available online at: http://www.fao.org/faostat/en/#data/QC (accessed January 6, 2021).

Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., et al. (2016). AlphaSim: software for breeding program simulation. *Plant Genome* 9, 1–14. doi: 10.3835/plantgenome2016.02.0013

Fox, C. M., Cary, T. R., Colgrove, A. L., Nafziger, E. D., Haudenshield, J. S., Hartman, G. L., et al. (2013). Estimating soybean genetic gain for yield in the Northern United States—Influence of cropping history. *Crop. Sci.* 53, 2473–2482. doi: 10.2135/cropsci2012.12.0687

Fu, Y.-B. (2015). Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet.* 128, 2131–2142. doi: 10.1007/s00122-015-2585-y

Gaynor, C., Gorjanc, G., Wilson, D., Hickey, J., and Money, D. (2020). *AlphaSimR: Breeding Program Simulations*. Available online at: https://CRAN.R-project.org/package=AlphaSimR (accessed July 6, 2020).

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., et al. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Sci.* 57, 2372–2386. doi: 10.2135/cropsci2016.09.0742

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P. S. (2017). Improving response in genomic selection with a population-based selection strategy: optimal population value selection. *Genetics* 206, 1675–1682. doi: 10.1534/genetics.116.197103

Gonen, S., Ros-Freixedes, R., Battagin, M., Gorjanc, G., and Hickey, J. M. (2017). A method for the allocation of sequencing resources in genotyped livestock populations. *Genet. Sel. Evol.* 49:47. doi: 10.1186/s12711-017-0322-5

Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolin, R., Gaynor, R. C., and Hickey, J. M. (2017a). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Sci.* 57, 216–228. doi: 10.2135/cropsci2016.06.0526

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolin, R., and Hickey, J. M. (2017b). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Sci.* 57, 1404–1420. doi: 10.2135/cropsci2016.08.0675

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

Gorjanc, G., and Hickey, J. M. (2018). AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. *Bioinformatics* 34, 3408–3411. doi: 10.1093/bioinformatics/bty375

Habier, D., and Fernando, R. L. (2009). Genomic selection using low-density marker panels. *Genetics* 182, 343–353. doi: 10.1534/genetics.108.100289

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. doi: 10.1186/1471-2105-12-186

Haun, W., Coffman, A., Clasen, B. M., Demorest, Z. L., Lowy, A., Ray, E., et al. (2014). Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant Biotechnol. J.* 12, 934–940. doi: 10.1111/pbi.12201

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690. doi: 10.2135/cropsci2009.11.0662

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* 54, 1476–1488. doi: 10.2135/cropsci2013.03.0195

Hickey, L. T., Hafeez, N. A., Robinson, H., Jackson, S. A., Leal-Bertioli, S. C. M., Tester, M., et al. (2019). Breeding crops to feed 10 billion. *Nat. Biotechnol.* 37, 744–754. doi: 10.1038/s41587-019-0152-9

Jacobs, T. B., LaFayette, P. R., Schmitz, R. J., and Parrott, W. A. (2015). Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnol.* 15:16. doi: 10.1186/s12896-015-0131-2

Jähne, F., Hahn, V., Würschum, T., and Leiser, W. L. (2020). Speed breeding short-day crops by LED-controlled light schemes. *Theor. Appl. Genet.* 133, 2335–2342. doi: 10.1007/s00122-020-03601-4

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42:35. doi: 10.1186/1297-9686-42-35

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., et al. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15:740. doi: 10.1186/1471-2164-15-740

Jarquin, D., Specht, J., and Lorenz, A. (2016). Prospects of genomic prediction in the USDA soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3* 6, 2329–2341. doi: 10.1534/g3.116.031443

Kemper, K. E., Bowman, P. J., Pryce, J. E., Hayes, B. J., and Goddard, M. E. (2012). Long-term selection strategies for complex traits using high-density genetic markers. *J. Dairy Sci.* 95, 4646–4656. doi: 10.3168/jds.2011-5289

Khan, H. (2013). A review on induced mutagenesis in soybean. *J. Cereals Oilseeds* 4, 19–25. doi: 10.5897/JCO10.004

Koester, R. P., Skoneczka, J. A., Cary, T. R., Diers, B. W., and Ainsworth, E. A. (2014). Historical gains in soybean (*Glycine max* Merr.) seed yield are driven by linear increases in light interception, energy conversion, and partitioning efficiencies. *J. Exp. Bot.* 65, 3311–3321. doi: 10.1093/jxb/eru187

Lange, C. E., and Federizzi, L. C. (2009). Estimation of soybean genetic progress in the South of Brazil using multi-environmental yield trials. *Sci. Agric.* 66, 309–316. doi: 10.1590/S0103-90162009000300005

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. *Front. Genet.* 9:237. doi: 10.3389/fgene.2018.00237

Lindgren, D., Wei, R.-P., and Lee, S. J. (1997). How to calculate optimum family number when starting a breeding program. *For. Sci.* 43, 206–212. doi: 10.1093/forestscience/43.2.206

Ma, Y., Reif, J. C., Jiang, Y., Wen, Z., Wang, D., Liu, Z., et al. (2016). Potential of marker selection to increase prediction accuracy of genomic selection in soybean (*Glycine max* L.). *Mol. Breed.* 36:113. doi: 10.1007/s11032-016-0504-9

Martin, S. K. S. (1982). Effective population size for the soybean improvement program in maturity groups 00 to IV1. *Crop Sci.* 22, 151–152. doi: 10.2135/cropsci1982.0011183X002200010035x

Matei, G., Woyann, L. G., Milioli, A. S., de Bem Oliveira, I., Zdziarski, A. D., Zanella, R., et al. (2018). Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. *Mol Breed.* 38:117. doi: 10.1007/s11032-018-0872-4

Meuwissen, T. H. E. (2009). Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* 41:35. doi: 10.1186/1297-9686-41-35

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

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

Muleta, K. T., Pressoir, G., and Morris, G. P. (2018). Optimizing genomic selection for a sorghum breeding program in haiti: a simulation study. *G3*, 9, 391–401. doi: 10.1534/g3.118.200932

Nagatoshi, Y., and Fujita, Y. (2019). Accelerating Soybean Breeding in a CO2-Supplemented Growth Chamber. *Plant Cell Physiol.* 60, 77–84. doi: 10.1093/pcp/pcy189

Nejati-Javaremi, A., Smith, C., and Gibson, J. P. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Anim. Sci.* 75, 1738–1745. doi: 10.2527/1997.7571738x

Ooi, H., Corporation, M., Weston, S., and Tenenbaum, D. (2019). *doParallel: Foreach Parallel Adaptor for the "parallel" Package (Version 1.0.16)*. Available online at: https://CRAN.R-project.org/package=doParallel (accessed July 2, 2020).

Paixão, T., and Barton, N. H. (2016). The effect of gene interactions on the long-term response to selection. *PNAS* 113, 4422–4427. doi: 10.1073/pnas.1518830113

Pook, T., Schlather, M., and Simianer, H. (2019). MoBPS - modular breeding program simulator. *bioRxiv* [Preprint]. doi: 10.1101/829333

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rogers, J., Chen, P., Shi, A., Zhang, B., Scaboo, A., Smith, S. F., et al. (2015). Agronomic performance and genetic progress of selected historical soybean varieties in the southern USA. *Plant Breed.* 134, 85–93. doi: 10.1111/pbr.12222

Ru, S., and Bernardo, R. (2019). Targeted recombination to increase genetic gain in self-pollinated species. *Theor. Appl. Genet.* 132, 289–300. doi: 10.1007/s00122-018-3216-1

Ru, S., and Bernardo, R. (2020). Predicted genetic gains from introgressing chromosome segments from exotic germplasm into an elite soybean cultivar. *Theor. Appl. Genet.* 133, 605–614. doi: 10.1007/s00122-019-03490-2

Santantonio, N., and Robbins, K. (2020). A hybrid optimal contribution approach to drive short-term gains while maintaining long-term sustainability in a modern plant breeding program. *bioRxiv* [Preprint]. doi: 10.1101/2020.01.08.899039

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Sebastian, S. A., Streit, L. G., Stephens, P. A., Thompson, J. A., Hedges, B. R., Fabrizius, M. A., et al. (2010). Context-specific marker-assisted selection for improved grain yield in elite soybean populations. *Crop Sci.* 50, 1196–1206. doi: 10.2135/cropsci2009.02.0078

Smallwood, C. J., Saxton, A. M., Gillman, J. D., Bhandari, H. S., Wadl, P. A., Fallen, B. D., et al. (2019). Context-specific genomic selection strategies outperform phenotypic selection for soybean quantitative traits in the progeny row stage. *Crop Sci.* 59, 54–67. doi: 10.2135/cropsci2018.03.0197

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An Improved Genotyping by Sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* 8:e54603. doi: 10.1371/journal.pone.0054603

Stewart-Brown, B. B., Song, Q., Vaughn, J. N., and Li, Z. (2019). Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3* 9, 2253–2265. doi: 10.1534/g3.118.200917

Sun, X., Hu, Z., Chen, R., Jiang, Q., Song, G., Zhang, H., et al. (2015). Targeted mutagenesis in soybean using the CRISPR-Cas9 system. *Sci. Rep.* 5:10342. doi: 10.1038/srep10342

Todeschini, M. H., Milioli, A. S., Rosa, A. C., Dallacorte, L. V., Panho, M. C., Marchese, J. A., et al. (2019). Soybean genetic progress in South Brazil: physiological, phenological and agronomic traits. *Euphytica* 215:124.

Toledo, F. H., Pérez-Rodríguez, P., Crossa, J., and Burgueño, J. (2019). isqg: a binary framework for in silico quantitative genetics. *G3* 9, 2425–2428. doi: 10.1534/g3.119.400373

Valliyodan, B., Ye, H., Song, L., Murphy, M., Shannon, J. G., and Nguyen, H. T. (2017). Genetic diversity and genomic strategies for improving drought and waterlogging tolerance in soybeans. *J. Exp. Bot.* 68, 1835–1849. doi: 10.1093/jxb/erw433

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

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., and Weigel, K. A. (2011). Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10. doi: 10.1186/1297-9686-43-10

Vaughn, J. N., and Li, Z. (2016). Genomic signatures of North American soybean improvement inform diversity enrichment strategies and clarify the impact of hybridization. *G3* 6, 2693–2705. doi: 10.1534/g3.116.029215

Wray, N., and Goddard, M. (1994). Increasing long-term response to selection. *Genet. Sel. Evol.* 26, 431–451. doi: 10.1186/1297-9686-26-5-431

Wright, M. N., Wager, S., and Probst, P. (2020). *ranger: A Fast Implementation of Random Forests*. Available online at: https://CRAN.R-project.org/package=ranger (accessed July 2, 2020).

Xavier, A. (2019). Efficient estimation of marker effects in plant breeding. *G3* 9, 3855–3866. doi: 10.1534/g3.119.400728

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., et al. (2018a). Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3* 8, 519–529. doi: 10.1534/g3.117.300300

Xavier, A., Muir, W. M., and Rainey, K. M. (2016). Assessing predictive properties of genome-wide selection in soybeans. *G3* 6, 2611–2616. doi: 10.1534/g3.116.032268

Xavier, A., Muir, W. M., and Rainey, K. M. (2019). bWGR: bayesian whole-genome regression. *Bioinformatics*. 36, 1957–1959. doi: 10.1093/bioinformatics/btz794

Xavier, A., and Rainey, K. M. (2020). Quantitative genomic dissection of soybean yield components. *G3* 10, 665–675. doi: 10.1534/g3.119.400896

Xavier, A., Thapa, R., Muir, W. M., and Rainey, K. M. (2018b). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet. Resour.* 16, 513–523. doi: 10.1017/S1479262118000102

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

Yu, J., Arbelbide, M., and Bernardo, R. (2005). Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. *Theor. Appl. Genet.* 110, 1061–1067. doi: 10.1007/s00122-005-1926-7

Zheng, N., Li, T., Dittman, J. D., Su, J., Li, R., Gassmann, W., et al. (2020). CRISPR/Cas9-based gene editing using egg cell-specific promoters in *Arabidopsis* and Soybean. *Front. Plant Sci.* 11:800. doi: 10.3389/fpls.2020.00800

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.