Journal of Crop Science and Biotechnology (2024) 27:349–358
https://doi.org/10.1007/s12892-024-00234-7

REVIEW ARTICLE

Balancing genomic selection efforts for allogamous plant breeding programs

Rafael Tassinari Resende[1,2]

Accepted: 14 January 2024 / Published online: 24 February 2024

Abstract

Genomic selection (GS) is fundamentally a statistical genetics technique, which encourages scientists to develop robust models for this purpose. However, the application of GS is not confined to mathematical theory alone; it entails a meticulous evaluation of its practicality and applicability, particularly across generations of crossbreeding and in the strategic management of base-populations used for model calibration. While costs have diminished, it remains a substantial investment, notably due to the dollar pricing of each breeding sample. To ensure the efficiency of this technology, foresight in planning is imperative, taking into account available data, those to be acquired, and the quality of SNP and phenotypic data. Maintaining focus on the base population that will endure throughout the selection cycles of the program is paramount (given that GS models are inherently linked to relatedness among individuals). Selection strategies encompassing both additive and non-additive effects are necessary. Still, they must be applied judiciously, considering the phase of the program, be it for the development of lines, hybrids, or both. The complexity of models should be managed with prudence, considering their routine applicability; for instance, a predictive artificial intelligence model may not always be the unequivocal choice. Furthermore, it is wise to consider that in some cases, a simple pedigree-based model may deliver results as effective and more cost-efficient than GS. However, when kinship information is limited or absent, this is where genomics reveals one of its greatest advantages. Genomic models possess a unique elegance, and those who employ them are at the forefront of crop biotechnology advancement.

Keywords Breeding populations · Genotyping · Recurrent selection · Precision breeding · Cross-validation · Phenomics

Introduction

Genomic selection (GS) represents a well-known innovation in the plant breeding process, allowing for the prediction of genotypic values without the need to grow and evaluate crops in the field. It proves particularly successful in recurrent selection programs tailored for allogamous species, characterized by elevated genomic heterozygosity rates and, in some cases, protracted selection cycles (Zhang et al. 2017a; Grattapaglia 2022). The efficacy of GS is closely tied to the Linkage Disequilibrium (LD) phenomenon, a pattern that affects loci across the genome. This type of non-random association suggests that the inheritance of loci may not strictly adhere to the pattern of independent Mendelian segregation across generations (Liu et al. 2015; Skelly et al. 2016). Thus, even a locus not directly linked to the expression of a phenotypic trait can provide a significant understanding of it. This fundamental feature of segregative genetics ushered in a new era of plant breeding, known as genomic analysis, enabling strides in developing more productive crop varieties.

GS heavily relies on the genetic relationships among individuals. When applying a model developed for a specific population to predict genotypic values in another unrelated population, the likelihood of success is expected to be significantly reduced (Ramstetter et al. 2017; Merrick et al. 2022). The base population represents the initial group of plants from which breeding endeavors originate. Therefore, thoughtfully choosing the base population is essential when applying GS approaches to segregative allogamous individuals (Labroo et al. 2021; Grattapaglia 2022). This careful

✉ Rafael Tassinari Resende
rafael.tassinari@ufg.br

[1] Agronomy Department, Plant Breeding Sector, Universidade Federal de Goiás (UFG), Goiânia, GO 74690-900, Brazil

[2] TheCROP, a Precision-Breeding Startup: Enviromics, Phenomics and Genomics, Goiânia, GO 74690-900, Brazil

choice serves as the cornerstone for the accuracy and efficacy of genomic prediction along the breeding cycles of the program, leading to more pertinent outcomes in the context of genetic improvement.

It is common in the literature to find studies that perform the training and validation of GS models within the same population (as will be discussed in the present Review and also been discussed by Taylor 2014, Merrick et al. 2022, and Berro et al. 2019). While this may seem a reasonable approach, given the optimization of models in a fixed founder population, many of these studies group training and validation populations in one or a few specific experiments, often utilizing samples from a genetic panel of the same generation (see Ferrão et al. 2017, Simiqueli & Resende 2020, and Simiqueli et al. 2023 for more details on this). In this context, it is important to consider the relevance of training and validating models across different generations, such as between parents and offspring, or even other combinations of relatedness (Simiqueli et al. 2023); and also, between different environments and types/phases of breeding populations (Resende et al. 2021). This practice not only ensures a more authentic validation but also strengthens the robustness and reliability of genomic predictions.

We are immersed in an era of advanced information systems and big data analysis, where the integration of information and the use of Artificial Intelligence (AI) prevail (Harfouche et al. 2019; Xu et al. 2022; Montesinos-López et al. 2021). However, despite the abundance of technologies and automation, certain processes have been neglected due to the pursuit of quick results. Indeed, GS cannot be considered a low-cost technique, and it is dependent on hard quantitative genetics skills. To calibrate the models, reliable phenotypic data and SNP markers are required. This may explain the hesitancy in adopting GS in plant breeding programs in some sectors, either due to its potential cost or the inapplicability of the model to operational reality (Wartha and Lorenz 2021). Therefore, progress is needed to seize adoption opportunities in both the short and long terms, with considerations about cost-based feasibility, disruption of current practices, and associated risks (Bernardo 2021).

Genomics demonstrates its great efficiency in the genetic characterization of individuals and populations (Huisman 2017; Dwiningsih et al. 2020). On the flip side, establishing DNA causality in the expression of complex phenotypes is a finicky field of study, susceptible to producing false positives in statistical models if not approached with scrupulous precision (Wu et al. 2018). The secret to the effective functioning of GS lies in the ability to unravel genetic relatedness among evaluated individuals through LD. It is worth noting that, in many cases, information about the pedigree of these individuals is already available. However, if reliable pedigree information is accessible for the target breeding population, the traditional approach of phenotypic selection not only proves equally effective but, in most cases, surpasses GS (Henryon et al. 2019; Michel et al. 2020). Selection-based exclusively on pedigree becomes a choice that can be adopted without hesitation. However, in scenarios where genealogical information is unavailable, or when seeking to maximize selection gains through the combination of pedigree and SNP data, the application of genomic selection techniques proves highly valid.

This brief review article offers planning perceptions into the field of Genomic Selection (GS), aiming to optimize time and resource investments during its implementation. It consistently centers around the driving force of the technique, the Linkage Disequilibrium (LD). The topics encompass high-throughput SNP-based genotyping of the breeding samples, GS design regarding sample sizes (both phenotyped-and-genotyped individuals), and predictive accuracies. The primary emphasis is on allogamous crops like maize (an annual crop) and eucalyptus (a perennial crop). The objective is not to compare predictive models or propose statistical protocols, but rather to aggregate strategies for the effective application of the technique and explore projections of predictive abilities based on available data.

SNP-based genotyping of the breeding samples

In the field of molecular genetics, the choice of the most suitable genetic marker is pivotal for precise and effective analysis. Studies indicate that single nucleotide polymorphisms (SNPs) can provide the most informative estimates of genetic differentiation and structure (Dwiningsih et al. 2020). These SNPs, being the most common form of DNA variation, allow for the simultaneous genotyping of hundreds to thousands of loci. Advances in sequencing technologies have contributed to the creation of extensive SNP datasets, expanding the viability of this approach. Today, it is easy to access public phenotypic and genomic data, for example, in wheat (an autogamous crop) and eucalyptus (Scheben et al. 2019; Resende et al. 2017). However, it is worth noting that some high-impact scientific journals request authors to share phenotypic and genomic data to facilitate publication.

In any genomic analysis, especially in the context of allogamous plant breeding, Linkage Disequilibrium (LD) plays the main role in relating DNA markers to agronomically relevant traits. LD represents the relationship between two or more loci along the genome, resulting in their dependent segregation (Skelly et al. 2016). Therefore, even a locus that may not be a direct expression of interest can provide relevant information if it is in LD with the target gene. This phenomenon is of great importance since, without the occurrence of LD, genomic technologies enabling rapid haplotype identification and SNP marker detection would be severely

compromised. Associating DNA with phenotypic traits would, in this context, be akin to the challenge of "looking for a needle in a haystack". Establishing the magnitude of LD is fundamental for conducting studies of assisted selection and genomic selection, although the identification of causal variants underlying phenotypic variation remains a considerable challenge.

It is not surprising that next-generation sequencing (NGS) technologies represent transformative genomic tools (Kim et al. 2020). Initiated between 2004 and 2005 with the Roche GS20 model, NGS platforms continue to evolve, enabling the generation of data with billions of base pairs and high nucleotide-level accuracy (Kchouk et al. 2017). The practical application of this process involves the fragmentation and preparation of DNA from plant samples, followed by the construction of genomic libraries. The NGS sequencer reads DNA sequences, and the resulting data are aligned to a species' reference genome, allowing for the identification of variants such as SNPs and indels.

Genotyping arrays, such as SNP chips, are widely used in plant breeding. Operationally, these chips consist of glass substrates containing thousands of oligonucleotides representing specific SNPs, enabling high-throughput genotyping. With high reproducibility and faster result analysis, SNP chips integrate molecular data from different studies, becoming a routine research approach (Rasheed et al. 2017). Figure 1A, adapted from Resende et al. 2017, provides a detailed illustration of the distribution of genotyped SNP markers on an Illumina BeadChip platform (Silva-Junior et al. 2015), specifically for a hybrid population of *Eucalyptus grandis* × *Eucalyptus urophylla*. Delving deeper into Fig. 1-B, we examine the patterns of allogamous heterozygosity across the genome, represented by $2pq = 2p(1-p)$, where 'p' signifies the frequency of one of the bialleles, and 'q' is the frequency of the other, or simply $1-p$. This illustration highlights the applicability of SNP chips in species with relatively complex genomes, such as eucalyptus, which has several native species in Oceania. For instance, *E. grandis* has a genome of approximately 697 Mb, while in *E. urophylla*, it is approximately 626 Mb.

To ensure high-quality genomic data, it is necessary to conduct data mining and cleaning. Genotyping platforms typically offer tens or hundreds of thousands of markers. However, in the sample used to calibrate the GS models, some marker polymorphisms may not be present. This implies that while the genotyping library identifies thousands of SNPs, only a portion of them may exhibit polymorphism in the sample. Therefore, we need to exclude markers with low or no frequency (Minor Allele Frequency, MAF), ensuring a high Call Rate (indicative of high-quality data), among other quality control measures. Please note that the sample in Fig. 1 has ~ 25 K markers, while the chip has 60 K (Silva-Junior et al. 2015). In addition, markers should be parameterized to capture both additive and non-additive genetic effects (Vitezica et al. 2013; Muñoz et al. 2014). Detailing these issues is not the focus of this text. For this purpose, the utilization of R packages like {snpReady} (Granato et al. 2018) and {AGHmatrix} (Amadeu et al. 2023) are very good options.

GS efforts planning: sample sizes and the predictive accuracy

Through high-throughput genotyping techniques, the availability of large quantities of SNPs per sample allows us to fully explore the genetic variance within breeding populations (Yang et al. 2017; Grattapaglia 2022). As previously mentioned, genomic selection is made possible by Linkage Disequilibrium (LD) among markers, which is, more directly put, the correlation that exists between two markers in a genotyped population with SNP markers. This means that

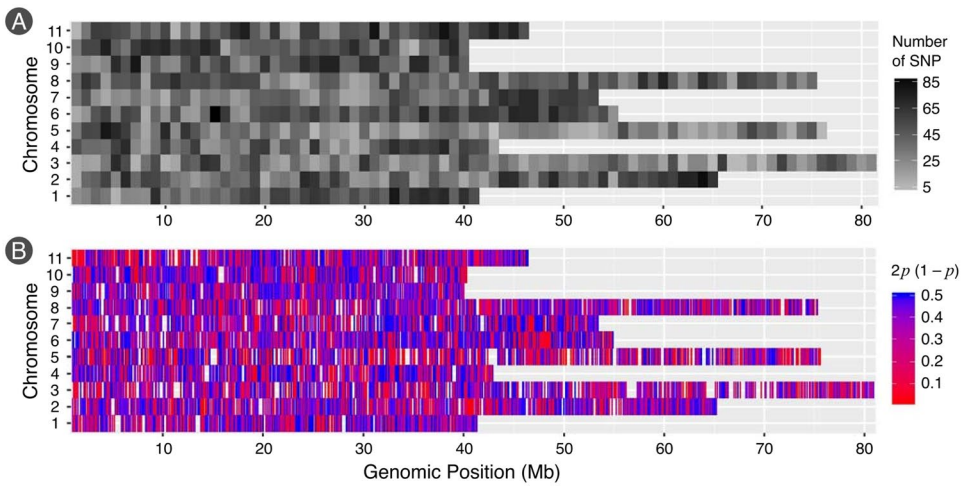

Fig. 1 Distribution of 24,806 polymorphic SNP markers along the 11 chromosomes of a eucalyptus population. Part "**A**" of the figure displays the concentration of SNPs per 1 Mb window. Part "**B**" of the figure shows the average heterozygosity of SNPs in a 100 kb window (in terms of the allele frequency 'p' and '$q = 1 - p$'). Adapted from Resende et al. (2017)

even if a marker is not directly linked to a QTL, many markers adjacent to the QTL can provide information about the segregation of genes related to the expression of the trait of interest (e.g., crop yield, time-growth, and biotic/abiotic stress tolerances).

By applying genomic-wide selection, or simply genomic selection (GS) techniques, it is possible to predict the "phenotypes" (i.e., genotypic values) of experimental breeding trials, even before these trials are conducted. In other words, based solely on the DNA of individuals who would hypothetically be planted in the field. Although this may initially seem unrealistic, it makes sense to remember that a significant portion of the information leading to the final phenotype originates from genes. By appropriately managing and/or correcting the environmental component of the phenotype (as is well known, Phenotype = Genotype + Environment), GS models will demonstrate strong predictive abilities (Montesinos-López et al. 2018).

It is important to note that, in terms of predictive ability (i.e., accurately ranking the best genetic materials), GS may not necessarily surpass the phenotypic selection carried out in the field (Heffner et al. 2011). This is because GS is generally applied as an "extremely-early-indirect" selection method, to approximate the direct selection (a "benchmark") conducted in the field (see preliminary Fig. 2A). However, GS can indeed be more advantageous than direct field selection, primarily for five reasons: *(i)* time savings, as early genomic selection can be conducted from initial plant propagules; *(ii)* resource savings, including labor and inputs that would be expended in the entire process of establishment, phenotypic measurement, and harvest/transportation in experimental breeding trials; *(iii)* the ability to evaluate a greater number of genetic materials that may not eventually go to the field and therefore would not be tested—for example, instead of taking 2000 genetic materials for field evaluation, 4000 could be genotyped and their phenotypes predicted genomically; *(iv)* predicting traits that are difficult to measure, such as root volume in cassava and wood volume in highly branched trees; and *(v)* correcting potential pedigree errors in the construction of the relationship matrix (*A*), with this information being recovered through the genomic matrix (*G*), or even concatenating matrices $A + G$ into a "super" matrix called "*H*" (Legarra et al. 2014).

The GS model performance can be understood with a didactic example involving a hundred genetic materials and a predictive ability of approximately 50% (Fig. 2). Some phenotypically good individuals may be left out of the genomic selection sieve and still demonstrate satisfactory predictive abilities. Figure 2B shows a shuffle of the best and worst phenotypes when predicted genomically, while Fig. 2C illustrates the relationship between observed phenotypic values (or genetic values estimated from means or *Best Linear Unbiased Prediction*—BLUP—of an experiment). It is observed that the term "selection" in GS can be confused with genomic "exclusion," which is actually what happens in most cases when GS is used to eliminate the worst individuals, rather than to effectively select the best.

Much is said about the appropriate number of markers to be used in a genomic selection process, as well as the ideal quantity of individuals/genetic materials to be phenotyped and genotyped (Merrick et al. 2022; Werner et al. 2020). The truth is that there is no one-size-fits-all approach to this process. It will depend on various factors inherent to the crop species, the population to be improved, the breeding

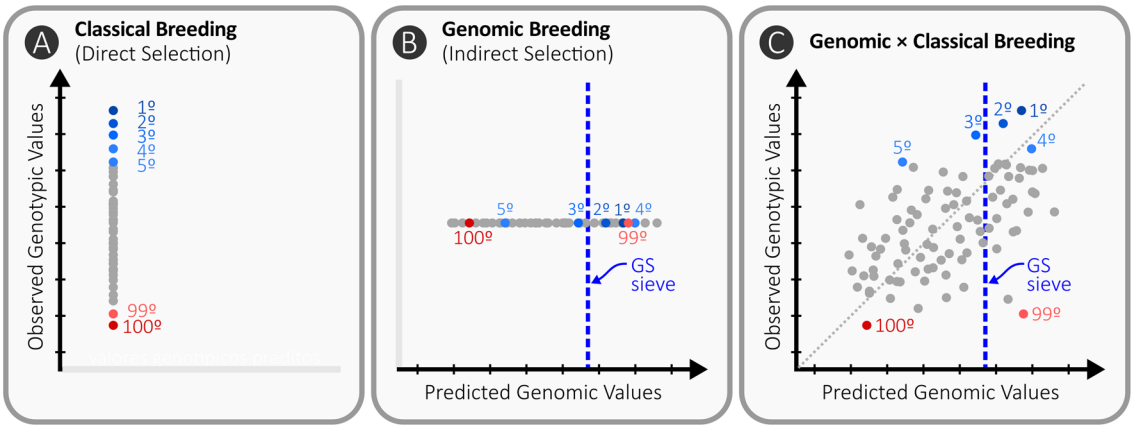

Fig. 2 Relationship between observed phenotypic values (averages or BLUP of hundred genetic materials) in the experiments versus genomically predicted values. In part "**A**," the so-called "true" genotypic ranking based on experimental field-measured values is shown. In part "**B**," only genomic prediction is displayed, yet analogously to the ranking of genetic materials seen in "**A**." Part "**C**" illustrates the relationship between parts "**A**" and "**B**." The dotted blue line represents the selection sieve of the GS, with selected genetic materials on the right and discarded ones on the left. The gray dotted line in part "**C**" provides an indication of the predictive ability of the GS model. Adapted from Vianello, Resende & Brondani (2023)

objective (such as developing lines, hybrids, and clones), and the phenotypic trait targeted for breeding (Silva et al. 2021). Among some equations used to plan the breeding program coupled with genomic selection, Resende (2008) proposed the equation:

$$\hat{r}_{g\hat{g}} = \sqrt{\left\{1 + \frac{4N_eL}{n_m} + \frac{2N_eL(4N_eL + n_m)^2}{Nh^2[\ln(2N_e)]n_m^2}\right\}^{-1}}$$

where $\hat{r}_{g\hat{g}}$ is the projected, or theoretical, predictive accuracy (attempting to anticipate the predictive ability achieved in GS practice); L is the size of the species genome (in morgans, M); n_m is the number of SNP-type markers; h^2 is the coefficient of heritability for the phenotypic trait (which can be broad-sense or narrow-sense heritability, depending on the breeding phase, we will delve further into this shortly); N is the actual size (i.e., sampled individuals) of the population (taking into account an equal number of individuals per family); N_e is the effective size of the population, representing the number of genetically contributing individuals to future generations, factoring in inbreeding effects that lower genetic diversity and can decrease N_e in smaller populations by increasing the likelihood of mating among close relatives. It is advisable to plan and have an understanding of the possible outcomes before starting any genomic selection process, as failing to do so may lead to significant resource and time losses. It is suggested to take a look at the alternatives to this equation found in Resende et al., (2012) at pages 139–146.

Therefore, the GS requires careful planning, with meticulous accounting of all resources to ensure its effectiveness. Figure 3 illustrates the application of the $\hat{r}_{g\hat{g}}$ equation proposed by Resende (2008). It highlights a genus with a relatively compact genome (*Eucalyptus spp.*, please referee to Fig. 1, wherein $L \approx 10$ M, with approx. 700 Mb length) (Bartholomé et al. 2015; Silva-Junior & Grattapaglia 2015) and another with a substantially larger genome (Zea mays, $L = 19.96$ M, with approx. 2,300 Mb length) (Dell'Acqua et al. 2015). In both cases, a feasible effective population size (N_e) of 50 was considered for crop improvement programs.

The $\hat{r}_{g\hat{g}}$ values presented at Fig. 3 assume that all events unfold as planned, with the minimization of any unforeseen contingencies not already accounted for in the calculation of the phenotypic trait heritability (h^2). Generally, genotyping technologies availability ranges from 1 K to over 600 K SNP markers for various crop species (Rasheed et al. 2017). Notice that, the quantity of utilized SNPs (n_m) plays a significant role in predictive capacity. It is also worth noting that, in general, GS demonstrates effectiveness for traits with both high and low h^2, with predictions being more accurate for traits with higher h^2. For traits with high h^2, the number of phenotyped-and-genotyped

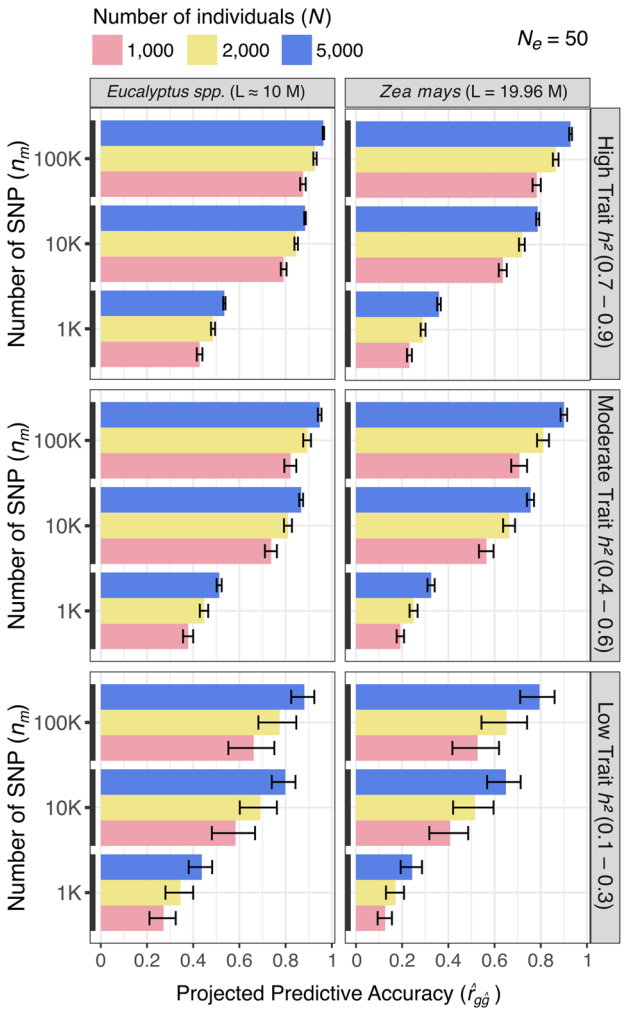

Fig. 3 Genomic selection projected scenarios using the $\hat{r}_{g\hat{g}}$ equation by Resende et al. (2012). It addresses genome sizes of *Eucalyptus spp.* ($L \approx 10$ Morgans, M) and *Zea mays* ($L = 19.96$ Morgans, M). The effective population size (N_e) is fixed in 50 for both. Predictive accuracies ($\hat{r}_{g\hat{g}}$) consider 1–100 K SNPs. Three scenarios of trait heritability (h^2 = {Low, Moderate, High}); and phenotyped-and-genotyped individuals (N = {1000, 2000, 5000}) were evaluated

individuals (N) is not a major limiting factor. Conversely, this is the case for traits with low h^2, where even populations of 5000 individuals may not provide substantial predictive accuracy.

At this point in the article, it is important to emphasize that the statistical-mathematical validation of broad GS models is necessary to assess the predictive ability of the models. Among the methods available for computing predictive abilities, the simple correlation between observed phenotypic values and values predicted by the model is a direct measure of the model's predictive reliability and is considered a trustworthy way to assess the performance of the GS model. Accuracy can also be calculated by weighting based on the heritabilities of the phenotypic and/or genomic

traits, with the premise of correcting for potential prediction shrinkage effects (Müller et al. 2015). However, caution must be exercised when including these quantities in accuracy equations, as this may lead to an overestimation of GS accuracy in traits with low heritability or an underestimation of predictive ability in traits with high heritability. To address this issue, a reasonable option is to use Pearson's correlation, a well-known measure of correlation, or even Spearman correlation (for genotypic rankings).

Regarding the sample size for GS, in general, it has been experienced that well over the stereotypically indicated 1000 individuals are needed to fit good GS models (see Fig. 3). Furthermore, after observing many efforts to fit various types/approaches of predictive models (such as Bayesian—Bayes A, B, Cπ, LASSO—and those via Artificial Intelligence or Machine Learning), there is little incremental gain in predictive ability over the initially described GBLUP or RRBLUP methods by Meuwissen et al. (2001). In fact, there are many situations where better fits can be achieved with more elaborate methods compared to the classic GBLUP/RRBLUP mentioned here (Montesinos-López et al. 2021). However, it is important to highlight those other efforts, such as managing breeding populations and strategies for feeding and validating the models, are vital to the success of GS. In addition, properly exploring the additive and non-additive components of the intended phenotypic traits, as well as adapting GS to the specific phase of the breeding program, are factors that generally provide greater benefits compared to striving for higher predictive abilities among Frequentist × Bayesian × AI methods.

Practical inferences on genomic selection in allogamous plant breeding

Numerous modeling approaches could be applied in genomic prediction, and Fig. 4 presents the basic mixed linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{e}$ as a single illustration. In this case, $\boldsymbol{y}$ is the vector of phenotypic data, $\boldsymbol{\beta}$ is the vector of fixed effects (such as: experimental replicates/blocks, locations, repeated measurements over time, among others); $\boldsymbol{g}$ is the vector of random genetic effects (the genetic materials, which can be lines, hybrids, among others); and $\boldsymbol{e}$ is the random vector of residuals; $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices on the fixed and random effects, respectively. It is not the focus of this article to discuss when to assign certain effects as fixed or random in nature, but it is a consensus that genetic materials should be considered as random in order to enable the execution of mixed models for genomic selection (Resende et al. 2008).

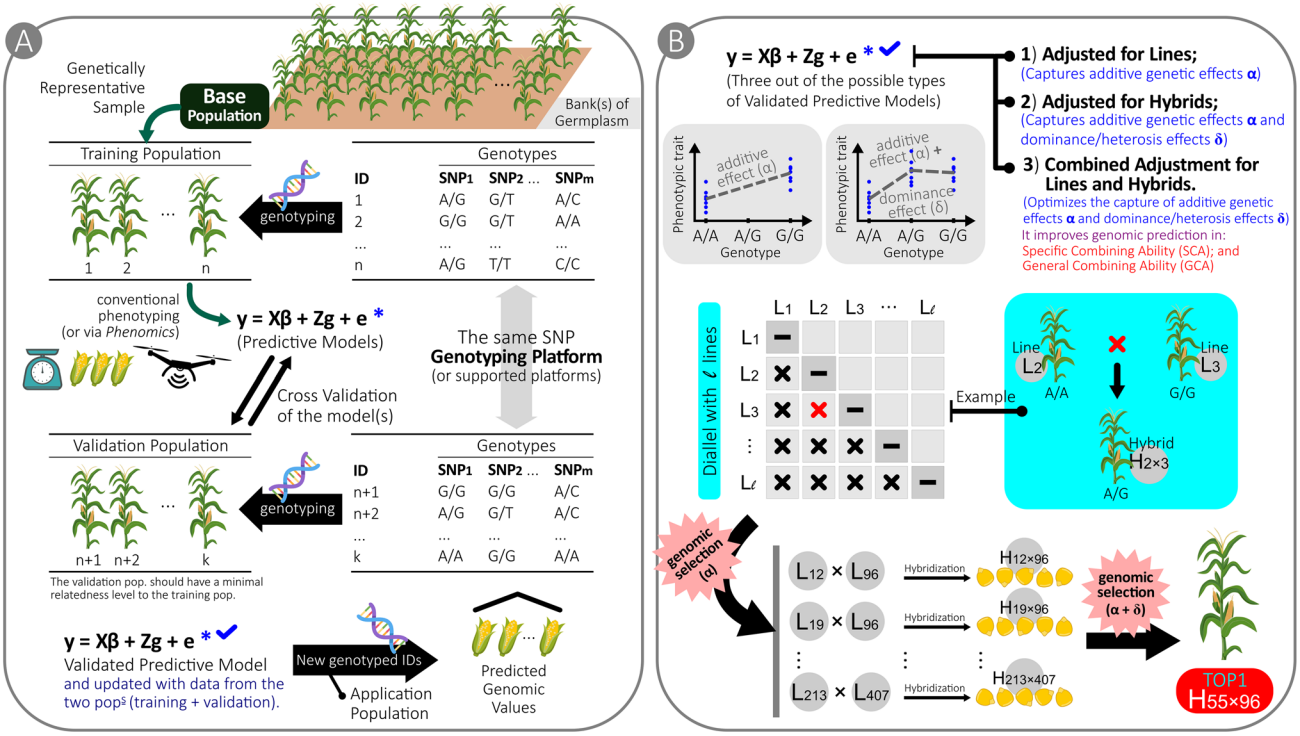

Fig. 4 Simplified diagram of a genomic selection (GS) process, using corn as an example. Part '**A**' of the figure illustrates the process of fitting and validating predictive genomic models. Part '**B**' of the figure depicts possible schemes for use, both for the prediction of inbred lines and for the prediction of hybrids based on the best inbred lines or validated models with hybrid information. Adapted from Vianello, Resende & Brondani (2023)

In the example procedure illustrated in Fig. 4, GS is employed for two purposes: first, in the development of improved inbred lines (L_1, L_2, …), and second, in the creation of improved heterotic hybrids ($H_{1\times2}$, $H_{1\times2}$, …). These genetic materials may eventually become registered cultivars, following appropriate field testing, such as in "Value for Cultivation and Use" (VCU) trials. It is important to note that a genomic predictive model will predict based on the data it has been trained on. If you feed it data from progeny tests, it will yield predicted values for maize progeny, just as it will for clonal eucalyptus tests if provided with clonal test data. Therefore, great care must be taken in selecting the base population for model training (Resende et al. 2022b).

The requirement for complete or partial genetic relatedness between training, validation, and application populations is a drawback of genomic selection. This is because LD, the driving force behind genomics, is easily lost and created between distinct populations, or even after several generations within the same population (Liu et al. 2015; Simiqueli et al. 2023). However, it is advisable to manage the model training data according to the program's objectives. Initially, it is need to comprehensively map the genetic base of the program (i.e., the germplasm bank), including the addition of new materials and, importantly, the removal of less desirable materials to eliminate unnecessary noise from the analysis to come (Resende et al. 2022b). It is worth noting that genetic materials developed using GS will always necessarily be related to the initial genetic base (Grattapaglia 2022). This is reasonable considering that companies typically have well-defined genetic bases. A good GS model is unlikely to predict the performance of genetic materials from other genetic bases (such as those from different companies, countries, or regions).

The same applies when using only one or a few environments in the genomic predictive model, as it will only be capable of predicting the performance of genetic materials for those few environments it has been trained on. The lack of representativeness in the input data for GS models, coupled with validation using a subset of the same data, can falsely inflate the perceived model performance, as the validation population will be entirely related to the training population. In addition, if the phenotypic data is collected in only one or a few environments, the model may not perform well in unobserved environments. Models that account for genotype–environment interactions (G × E) have been shown to be more effective than the traditional GBLUP model in terms of prediction accuracy (Montesinos-López et al. 2018). Two strategies can be employed to address this issue: *(i)* develop a multi-environment genomic model capable of providing predictions with high stability, meaning the genetic material's expected value is good regardless of the environment; *(ii)* use models that incorporate Genotype × Environment × Management interactions (G × E × M)—notable among these are models within the scope of Enviromics (Resende et al. 2021, 2022a; Costa-Neto et al. 2023), which can provide predictions for individuals with high stability and adaptability across different locations. These models can predict improved materials on a site-specific scale.

In this context, it is essential to select the key traits for operational or industrial model feeding. Some traits are easier to measure than others, but this can lead to a low genetic correlation with the actual trait of interest, a serious problem that is often overlooked. For example, phenotypic traits from genetic breeding tests (e.g., progeny, hybrids, clones, among others) may not correlate well with actual field performance. While it is important to feed the GS model with operational data, there is often limited data available for commercial genotypes. In such cases, integrating test and commercial data can help compute the genetic correlation between the two types of data. This approach can effectively address the issue and yield better results in grain production, forestry, horticulture, fruit farming, and other sectors (Resende et al. 2021).

Genomic selection for multiple traits is an approach in which plant breeders make selection decisions considering a variety of trait characteristics, such as yield, plant height, flowering time, and disease resistance, aiming to optimize genetic gain over multiple generations. However, while index selection is a common practice, it presents challenges in optimizing non-linear breeding objectives and in experimenting to determine the ideal weights for each trait (Moeinizade et al. 2020). Thus, in any breeding program incorporating genomic selection, various types of phenotypic traits will be improved, preferably simultaneously. Large-scale phenotyping data can also be included in GS models, such as data collected by sensors on drones or those predicted via Near Infrared Spectroscopy (NIRS) (Robert et al. 2022). Each phenotypic trait has its own characteristics, as well as genetic nature, inheritance, and so forth. The traits will be directly related to the type of genetic material being worked on, as well as the program's objectives. However, one thing is certain: the higher the heritabilities of the trait (whether additive in the narrow sense—h_a^2, or broad-sense heritability—h^2), the better the GS model will work and deliver good results (see Fig. 3, where the theoretical results will be approximate for both h^2 and h_a^2), i.e., in a recurrent selection program, in the initial phases, genetic variability tends to be greater, and due to the evaluation of many materials, likely with low replication, heritabilities tend to be higher (Zhang et al. 2017b). The paradox is that later in the program, although the genetic base may narrow (after several cycles of selection), the quantity of genetic materials is much greater and, therefore, they are more experimentally replicated.

Therefore, GS models should be managed considering that heritabilities should be maximized as much as possible to optimize selection efficiency. This is generally achieved in two ways: *(i)* by increasing genetic variation; and/or *(ii)* by reducing environmental variance through better residual control or increasing the number of replications. In the early stages of the program, whether it is autogamous or allogamous, annual or perennial, additive effects (α) are valued, as these involve crossing, recombination, and selection stages (non-additive effects tend to be less influential in these stages). However, in the later stages of the programs, non-additive effects (δ) are also desired, since the generated cultivars are usually hybrids with some degree of heterosis (a dominance phenomenon) and genetically more uniform materials (Labroo et al. 2021). While inbreeding is a desirable, even indispensable, process in the stages of composing pure lines, if not well managed, its consequent inbreeding depression can pose a significant problem during the hybridization stages, particularly in F1-segregating (outcrossing) crosses in perennial species, such as forest and fruit-bearing species. Its critical impact includes decreasing genetic diversity and vitality—a reduction in variability that is the opposite of heterosis—increasing susceptibility to diseases, and reducing reproductive success and overall population fitness.

In this regard, using Fig. 4B as a reference, one can employ a model for additive genomic prediction, which will be effective in predicting segregating individuals in the early stages of the program, even if the goal is to obtain pure lines. It is also possible to use models that incorporate additivity and dominance when considering only the materials at the end of the program. A modern approach involves integrating data from all stages, maximizing and interconnecting the entire selection process. But, it is advised that when adopting AI models for GS to capture pure additive effects, caution is necessary. GS AI-based models may inherently capture non-additive effects (i.e., non-linear) unless their architectures are ultra-simplified, but in such cases, traditional linear models can provide the similar results.

The combination of different sources of phenotypic data offers several advantages, starting with the use of various environments, which allows for predicting behavior in terms of stability and adaptability of genotypes. Furthermore, since it involves the same genetic base, the effective population size ($\boldsymbol{N_e}$) usually does not change drastically, but the total population size ($\boldsymbol{N}$) increases, allowing for the best of both worlds: high variability in the initial populations of the program and a greater number of experimental repetitions in the final stages. This has a direct impact on increasing the predictive capacity of the model.

The incorporation of multi-omic approaches, often combined with genomics, is also a powerful tool in the genetic prediction of plants. These approaches can integrate transcriptomic, proteomic, metabolomic, phenomics, genomic, and other omic data to predict phenotypes of the plant individuals under study. Aggregating information at a higher level of intimacy with the final phenotype can improve the predictive capacity of the models, as demonstrated in the use of exomic markers, which effectively translate into phenotypic proteins (Hashmi et al. 2015). With other molecular structures, the logic is similar. For example, the prediction of flavor compounds (sugars, acids, and volatiles) in blueberries and tomatoes, based on metabolomics, shows very promising results (Colantonio et al. 2022), as does genomics in coffee taste (Ferrão et al. 2023). Approaches that combine transcriptomics, proteomics, metabolomics, and functional genomics were also conducted for the study of abiotic stress in vegetables (Zhuang et al. 2014). Performing "genomic" prediction but with phenomics data (using NIR) is also something astonishing (Robert et al. 2022). The joint analysis of these different layers of information increases the predictive accuracies of the models and enables the prediction of characteristics highly influenced by the environment or even subjective, such as the taste of agricultural products.

Final considerations

Advancements in allogamous plant genomics are characterized by the ongoing evolution of molecular tools and the gradual yet significant reduction in genotyping costs. This trend widens the applicability of these technologies to an even broader range of samples within genetic improvement programs. While we observe a leaning towards the use of an increasing number of genomic markers in breeding analyses, potentially involving hundreds of thousands of SNP markers, there are also researchers advocating for a framework with a lower number of markers, however, supplemented with advanced data imputation techniques.

Genomics also plays a role in establishing more sustainable and cost-effective production processes, especially through the development of cultivars with increased disease resistance, particularly when these diseases follow an oligogenic pattern of expression. In addition, the growing integration of artificial intelligence and machine learning techniques in genomic analysis aims to automate and optimize the interpretation of extensive datasets.

There is also a growing trend in incorporating large-scale phenotyping data, such as those obtained through drones and NIRS, into genomic selection models. Alternatively, the use of analog models to genomics, also for predictive purposes, but with data (characterizing similarities between samples) derived from other omics, is being considered. Furthermore, the integration of envirotypic or enviromics data is also emerging as a rising practice, to better address variations stemming from genotype–environment interactions.

The GS is a statistical technique for crop breeding. Its application requires careful planning and pragmatic evaluation, especially across generations and at the base-population model management. Selection strategies should take into account the program phase and model complexity. In some cases, a pedigree-based model can be as effective as GS, but genomics excels when kinship information is limited. Undeniably, genomic models represent an elegant and advanced tool in crop breeding.

Acknowledgements To the Postgraduate Program in Genetics and Plant Breeding (PPGGMP) at the University of Goiás (UFG), to the Precision Breeding Laboratory—LAMP at UFG; to Dr. Rosana Vianello (Embrapa Rice and Beans), Dr. Cláudio Brondani (Embrapa Rice and Beans) and Dr. Leandro Neves (Rapid Genomics) for the clarifying suggestions in refining the text. Special thanks also are extended to Dr. Dario Grattapaglia (Embrapa Genetic Resources and Biotechnology) for support in the complexities of species genome length. This article was designed to support the postgraduate course in *Genomic Statistics* of the PPGGMP/UFG program.

Funding The author received no specific funding for this work.

Data availability There is no data to be made available in this review article.

Declarations

Conflict of interest There is no conflict of interest.

References

Amadeu RR, Garcia AAF, Munoz PR, Ferrão LFV (2023) AGHmatrix: genetic relationship matrices in R. Bioinformatics 39(7):445

Bartholomé J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Gion JM (2015) High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. New Phytol 206(4):1283–1296

Bernardo R (2021) Upgrading a maize breeding program via two-cycle genomewide selection: same cost, same or less time, and larger gains. Crop Sci 61(4):2444–2455

Berro I, Lado B, Nalin RS, Quincke M, Gutiérrez L (2019) Training population optimization for genomic selection. Plant Genome 12(3):190028

Colantonio V, Ferrão LFV, Tieman DM, Bliznyuk N, Sims C, Klee HJ, Resende MF Jr (2022) Metabolomic selection for enhanced fruit flavor. Proc Natl Acad Sci 119(7):e2115865119

Costa-Neto G, Crespo-Herrera L, Fradgley N, Gardner K, Bentley AR, Dreisigacker S, Crossa J (2023) Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. G3 13(2):313

Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, Pè ME (2015) Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Geno Biol 16(1):1–23

Dwiningsih Y, Rahmaningsih M, Alkahtani J (2020) Development of single nucleotide polymorphism (SNP) markers in tropical crops. Adv Sustaine Sci Eng Tech 2(2):343558

Ferrão LFV, Ortiz R, Garcia AF (2017) Genomic selection: state of the art. In: Genetic improvement of tropical crops. Springer, Cham. https://doi.org/10.1007/978-3-319-59819-2_2

Ferrão LFV, Dhakal R, Dias R, Tieman D, Whitaker V, Gore MA, Resende MF Jr (2023) Machine learning applications to improve flavor and nutritional content of horticultural crops through breeding and genetics. Curr Opin Biotechnol 83:102968

Granato IS, Galli G, de Oliveira Couto EG, Souza MBE, Mendonça LF, Fritsche-Neto R (2018) snpReady: a tool to assist breeders in genomic analysis. Mol Breed 38:1–7

Grattapaglia D (2022) Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. Forests 13(10):1554

Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Mugnozza GS, Altman A (2019) Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends Biotechnol 37(11):1217–1235

Hashmi U, Shafqat S, Khan F, Majid M, Hussain H, Kazi AG, Ahmad P (2015) Plant exomics: concepts, applications and methodologies in crop improvement. Plant Signal Behav 10(1):e976152

Heffner EL, Jannink JL, Iwata H, Souza E, Sorrells ME (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51(6):2597–2606

Henryon M, Liu H, Berg P, Su G, Nielsen HM, Gebregiwergis GT, Sørensen AC (2019) Pedigree relationships to control inbreeding in optimum-contribution selection realise more genetic gain than genomic relationships. Genet Sel Evol 51(1):1–12

Huisman J (2017) Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. Mol Ecol Resour 17(5):1009–1024

Kchouk M, Gibrat JF, Elloumi M (2017) Generations of sequencing technologies: from first to next generation. Biol Med 9(3):1–8

Kim KD, Kang Y, Kim C (2020) Application of genomic big data in plant breeding: past, present, and future. Plants 9(11):1454

Labroo MR, Studer AJ, Rutkoski JE (2021) Heterosis and hybrid crop breeding: a multidisciplinary review. Front Genet 12:643761

Legarra A, Christensen OF, Aguilar I, Misztal I (2014) Single step, a general approach for genomic selection. Livest Sci 166:54–65

Liu H, Zhou H, Wu Y, Li X, Zhao J, Zuo T, Pan G (2015) The impact of genetic relationship and linkage disequilibrium on genomic selection. PLoS ONE 10(7):e0132379

Merrick LF, Herr AW, Sandhu KS, Lozada DN, Carter AH (2022) Optimizing plant breeding programs for genomic selection. Agronomy 12(3):714

Meuwissen TH, Hayes BJ, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Michel S, Löschenberger F, Sparry E, Ametz C, Bürstmayr H (2020) Multi-year dynamics of single-step genomic prediction in an applied wheat breeding program. Agronomy 10(10):1591

Moeinizade S, Kusmec A, Hu G, Wang L, Schnable PS (2020) Multi-trait genomic selection methods for crop improvement. Genetics 215(4):931–945

Montesinos-López A, Montesinos-López OA, Gianola D, Crossa J, Hernández-Suárez CM (2018) Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3 Genes Genomes Genetics 8(12):3813–3828

Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JW, Fajardo-Flores SB, Crossa J (2021) A review of deep learning applications for genomic selection. BMC Geno 22:1–23

Müller D, Technow F, Melchinger AE (2015) Shrinkage estimation of the genomic relationship matrix can improve genomic estimated breeding values in the training set. Theor Appl Genet 128:693–703

Muñoz PR, Resende MF Jr, Gezan SA, Resende MDV, de Los Campos G, Kirst M, Peter GF (2014) Unraveling additive from non-additive effects using genomic relationship matrices. Genetics 198(4):1759–1768

Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, Williams AL (2017) Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. Genetics 207(1):75–82

Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z (2017) Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant 10(8):1047–1064

Resende MDV (2008) Genômica quantitativa e seleção no melhoramento de plantas perenes e animais. Embrapa Florestas, Colombo, p 330

Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D (2017) Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. Heredity 119(4):245–255

Resende RT, Piepho HP, Rosa GJ, Silva-Junior OBE, Silva FF, de Resende MDV, Grattapaglia D (2021) Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor Appl Genet 134:95–112

Resende RT, Chenu K, Rasmussen SK, Heinemann AB, Fritsche-Neto R (2022a) EDITORIAL: enviromics in plant breeding. Front Plant Sci 13:935380

Resende MPM, Filho AJC, Antunes AM, de Oliveira BM, de Oliveira RG (2022b) Population genomics of maize. In: Population genomics. Springer, Cham. https://doi.org/10.1007/13836_2022_101

Resende MDV, Silva FE, Lopes PS, Azevedo CF (2012) Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial. *Viçosa: Ed. UFV*

Robert P, Brault C, Rincent R, Segura V (2022) Phenomic selection: a new and efficient alternative to genomic selection genomic selection (GS). In: Bartholome J, Ahmadi N (eds) Genomic prediction of complex traits: methods and protocols. Springer, New York

Scheben A, Verpaalen B, Lawley CT, Chan CKK, Bayer PE, Batley J, Edwards D (2019) CropSNPdb: a database of SNP array data for brassica crops and hexaploid bread wheat. Plant J 98(1):142–152

Silva LA, Peixoto MA, Peixoto LDA, Romero JV, Bhering LL (2021) Multi-trait genomic selection indexes applied to identification of superior genotypes. Bragantia (80):e3621

Silva-Junior OB, Grattapaglia D (2015) Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New Phytol 208(3):830–845

Simiqueli GF, de Resende MDV (2020) Entropy and mutual information in genome-wide selection: the splitting of k-fold cross-validation sets and implications for tree breeding. Tree Genet Genomes 16:1–14

Simiqueli GF, Resende RT, Takahashi EK, de Souza JE, Grattapaglia D (2023) Realized genomic selection across generations in a reciprocal recurrent selection breeding program of *Eucalyptus* hybrids. Front Plant Sci 14:1252504

Skelly DA, Magwene PM, Stone EA (2016) Sporadic, global linkage disequilibrium between unlinked segregating sites. Genetics 202(2):427–437

Taylor JF (2014) Implementation and accuracy of genomic selection. Aquaculture 420:S8–S14

Vianello RP, Resende RT, Brondani C (2023) *Genômica*. In: Melhoramento de Precisão. Embrapa

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195(4):1223–1230

Wartha CA, Lorenz AJ (2021) Implementation of genomic selection in public-sector plant breeding programs: current status and opportunities. Crop Breed Appl Biotechnol 21(S):e394621S15

Werner CR, Gaynor RC, Gorjanc G, Hickey JM, Kox T, Abbadi A, Stahl A (2020) How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. Front Plant Sci 11:592977

Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Yang J (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun 9(1):918

Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Qian Q (2022) Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant 15(11):1664–1695

Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. Nat Genet 49(9):1304–1310

Zhang A, Wang H, Beyene Y, Semagn K, Liu Y, Cao S, Zhang X (2017a) Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. Front Plant Sci 8:1916

Zhang X, Pérez-Rodríguez P, Burgueño J, Olsen M, Buckler E, Atlin G, Crossa J (2017b) Rapid cycling genomic selection in a multiparental tropical maize population. G3: Genes Genomes Genetics 7(7):2315–2326

Zhuang J, Zhang J, Hou XL, Wang F, Xiong AS (2014) Transcriptomic, proteomic, metabolomic and functional genomic approaches for the study of abiotic stress in vegetable crops. Crit Rev Plant Sci 33(2–3):225–237