



Volume: 01 Issue No: 05

Accelerating Genetic gain through Genomic Selection in Crop Breeding

Article ID: AP-V01-I05-06

**Virendra Jangid^{1*}, Sree Harshita Sindam¹, Gopal Singh¹, Mohit Kumar Sharma¹
and Kumar Sanu¹**

¹ Department of Genetics and Plant Breeding,

Rajmata Vijayaraje Scindia Krishi Vishwavidyalaya, Gwalior, Madhya Pradesh, India

Corresponding Author's Email: jangidvirendra65@gmail.com

Abstract

Plant breeding has played a key role in agricultural development, helping to improve crop yield, disease resistance, and adaptability to changing environments. Traditionally, breeders mainly depended on phenotypic selection and Mendelian genetics, but these methods were often time-consuming and labour intensive. MAS is less successful for quantitative variables governed by numerous small-effect QTLs, but it is helpful for choosing genes or QTLs with significant effects.

MARS does this better by utilizing both phenotypic and marker data, but it still mostly relies on markers that are strongly linked to the characteristic. In order to get around this restriction, Meuwissen et al. (2001) suggested Genomic Selection (GS), which uses genome-wide markers, including markers that don't

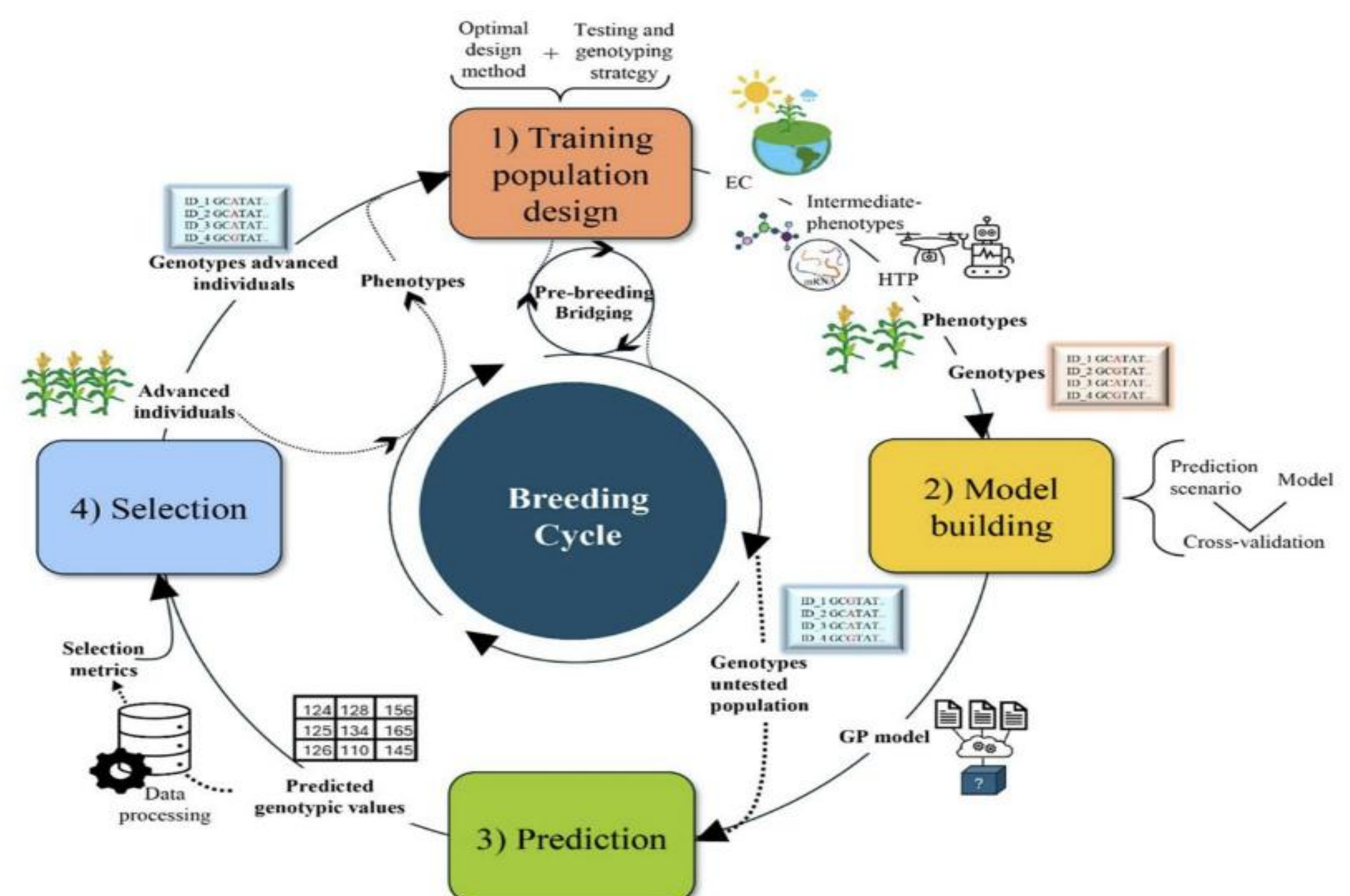
significantly correlate with the trait on their own. Plant breeding has been revolutionized by Genomic Selection (GS), which makes it possible to quickly and accurately identify superior plants. It can forecast how well plants will perform in key characteristics like yield, disease resistance, and stress tolerance. GS lowers breeding expenses, improves selection accuracy, and shortens the breeding cycle. It is extensively utilized in horticulture crops, perennial species, wheat, maize, rice, and other crops.

Genomic Selection (GS) is projected to become a widely used breeding approach for many crops as genomic technology become more affordable and computational techniques continue to progress.

Key words: genomic selection, QTL, marker, Breeding value

1. Introduction:

An enhanced version of MAS called Genomic Selection (GS) uses markers covering the complete genome. Each individual's Genomic Estimated Breeding Value (GEBV) is determined by estimating the consequences of all markers, including those with minor effects. Higher GEBV individuals are chosen for breeding. The genotypic value incorporates both additive



and non-additive genetic effects, whereas the breeding value (BV) indicates the expected performance of progeny and mostly reflects additive genetic effects. In GS, the breeding population is genotyped to predict GEBVs without thorough phenotyping, whereas a training population is genotyped and phenotype to create the prediction model. It is referred to as gene-assisted genomic selection when information on known QTLs is also incorporated.

GS can be used to choose parents and promising plants at various phases of the breeding cycle.

1. Training population
2. Model building
3. Prediction
4. Selection

2. Genomic Selection Process:

GS primarily requires two population a training population and a breeding population. To begin, an appropriate training population is chosen, and its plants are genotyped with genome-wide markers and phenotyped for the target traits.

This information is then used to build and train the GS model and to estimate marker effects. After that, the breeding population is genotyped with the same markers, although phenotyping is not needed. Using the marker data, the model computes GEBVs (Genomic Estimated Breeding Values) for every individual. In the end, plants with high GEBVs are chosen as superior lines or parents for subsequent breeding.

2.1. Training population:

In genomic selection (GS), the training population consists of a set of individuals or lines that possess both genotypic (marker) and phenotypic data, and these are employed to train the genomic prediction model. The training population ought to be representative of the breeding population and should adequately capture its genetic diversity as well as trait variation.

Random sampling is frequently employed, yet a random sample might not always be representative and can therefore result in lower prediction accuracy. A suitable training population should include individuals with divergent GEBVs, low marker collinearity, sufficient genetic diversity, and adequate recombination. Preferably, it should be composed of individuals that are closely related to, or share common ancestors with, the breeding population, since differences in allele frequencies, genetic background, QTL effects and linkage disequilibrium (LD) can lower GEBV accuracy.

The training population can be targeted, in which information about the target population guides selection, or untargeted, in which no such information is used. A range of optimization methods can be used to select the most useful individuals on the basis of genetic diversity, genetic relationships, PEV, CD and prediction accuracy. A training population may be made up of historical data or newly generated populations such as biparental crosses, doubled haploid populations, testcrosses or intermated lines. Ideally, a separate training population can be created for each breeding population, which yields high accuracy but raises phenotyping cost and time. As an alternative, a single large training population containing representative individuals from several breeding populations can be employed, which lowers cost and shortens the breeding cycle. The training population should also be updated on a regular basis, since recombination, selection and genetic drift alter allele frequencies and the LD relationship between markers and QTLs across generations.

When the training and breeding populations differ in their LD structures, prediction accuracy declines. Consequently, individuals from recent breeding cycles should be added and the GS model should be retrained. Historical phenotypic and genotypic data can be used as well, particularly when the data are extensive, have high heritability, and adequately represent the current genotypes and environments. Nevertheless, historical data may be unbalanced across years and influenced by genotype $\times$ environment interaction. Generally, GEBV prediction accuracy increases as training population size increases. Marker density needs to be high enough to guarantee that the majority of QTLs governing the target trait are in

strong linkage disequilibrium with at least one marker. The marker density required is influenced by the extent of LD, the crop species, trait heritability, and genetic diversity.

2.2. Model building:

To increase prediction accuracy by capturing phenotypic variance, a variety of genomic prediction (GP) models have been created. Parametric linear models, semi-parametric models, and non-parametric machine-learning models are the main methods. Stepwise regression, ridge regression (RR-BLUP), GBLUP, and Bayesian models like BayesA, BayesB, BayesC, Bayesian ridge regression, and Bayesian LASSO are examples of parametric models.

The primary assumption of these models is additive marker effects. Ridge regression is appropriate for traits controlled by numerous QTLs with minor effects since it assesses all marker effects concurrently and decreases them toward zero. Bayesian models can function effectively with high marker density and few phenotypic records, and they let marker effects and variances to vary. Complex and higher-order epistatic interactions can be explained using semi-parametric models, especially neural networks and reproducing kernel Hilbert space (RKHS). While neural networks are adaptable but computationally costly and may overfit, RKHS can capture complicated connections but rely on the right kernel selection. Large datasets can be handled by non-parametric machine-learning models like random forest (RF) and support vector machines (SVM), which are helpful when epistasis significantly contributes to genetic variance. Additionally, GP can incorporate biological information through the use of haplotype blocks or distinct marker classes, locus-specific or class-specific priors, and grouping markers based on gene ontology, metabolic pathways, or genome annotations. These methods can increase the accuracy of predictions.

Due to their general performance and ease of implementation, mixed and Bayesian linear models are employed extensively. Nonetheless, complicated features incorporating epistasis may benefit from semi-parametric and non-parametric models. As a result, several models may be chosen based on the genetic architecture, marker density, data type, and trait.

2.3. Prediction and selection:

Predicting total genotypic value is more difficult since it incorporates additive, dominance, and epistatic effects. Following prediction, selection should preserve genetic diversity while increasing genetic gain. Selection techniques that balance genetic gain, variety, and long-term breeding progress include weighted GS, anticipated cross value, utility criterion, ideal haploid/population value, look-ahead selection, and genomic mating.

3. Applications of GS in plant breeding:

Selection in conventional breeding is determined by phenotypic performance. To assess marker effects, a training population is genotyped and phenotype in genomic selection (GS). Individuals in the breeding

population are then selected based on their GEBVs, which are predicted using these effects. GEBVs can also be used to choose parents for hybridization. Therefore, superior lines from the breeding population are chosen for variety creation, and phenotyping is primarily restricted to the training population.

Creating recombinant inbred lines (RILs) from a hybrid between two homozygous lines is the first step in a straightforward GS-based recurrent selection strategy for self-pollinated crops. To train the GS model, the RILs are genotyped and assessed over a period of one to two years in replicated multilocation trials. Phenotypic performance is used to select superior RILs, which are then crossed. GS is applied to the resultant F_2 plants, and those with high GEBVs are chosen and crossed once more. After multiple iterations of this cycle, homozygous superior lines are isolated.

GS finds and chooses beneficial genetic diversity from gene banks for pre-breeding using genotyping and training populations. It preserves genetic diversity while introducing advantageous foreign alleles into superior breeding stock. GS helps bring advantageous alleles into breeding programs by accelerating the improvement of different donor lines and their crossings with elite lines.

GS requires phenotyping, and high-throughput phenotyping (HTP) makes it possible to assess numerous attributes quickly, accurately, and affordably utilizing sensors and images. While HTP traits can be utilized in multi-trait GS to enhance prediction, it can also increase genetic gain and selection accuracy. Random regression models (RRMs) can be used to assess longitudinal traits measured throughout crop growth. These models incorporate genetic changes over time and can increase prediction accuracy when compared to single-time-point models.

In GP, transcriptomic, proteomic, and metabolomic data can be combined with genomic data to enhance prediction accuracy and offer more biological information. Although tissue, sample time, data size, and ambient factors all affect accuracy, metabolomic data frequently provide helpful supplementary information. The application of multi-omics models, such as GBLUP, RRBLUP, Bayesian, SVM, RKHS, and RF, is constrained by their high cost, processing demand, and difficulties with data imputation.

GP first employed average phenotypic data or single-environment data; however, this eliminates some environmental information. Subsequently, GEI was modelled using AMMI, linear mixed models, factor-analytic models, joint regression, and factorial regression (FR). Weather and soil characteristics can be integrated into GP through envirotyping. Reaction norm and factor-analytic models capture genotype responses across environments, whereas GBLUP $M \times E$ models marker $\times$ environment interactions. To predict genotype performance in novel contexts, environmental indices (EI) like CERIS can be used in conjunction with GS via JGRA to detect significant environmental parameters. In general, GEI techniques strike a compromise between computation, interpretability, model complexity, and prediction accuracy.

environmental, and management characteristics to simulate crop growth. Because physiological traits need to be assessed across contexts, their calibration for novel genotypes is expensive and time-consuming. CGMs are combined with GS (CGM-GP) to get around this. While direct CGM-GP integration can enhance prediction in untested situations and specific genotype $\times$ management combinations, CGMs can produce stress variables from meteorological data. In addition to helping with biological inference and finding genetic loci that influence genotype-specific parameters, CGM-GP has been used in rice, wheat, and maize. Computational demand is still a significant barrier, though.

4. Future Potential of Genomic Selection:

The integration of genomic selection (GS) with cutting-edge technologies that boost breeding efficiency, speed up genetic gain, and improve prediction accuracy will determine the future of GS in plant breeding. Conventional breeding is anticipated to become predictive, data-driven, and precision-based through the integration of GS with high-throughput phenotyping (HTP), multi-omics, artificial intelligence (AI), remote sensing, gene editing, and speed breeding. The production of high-yielding, climate-resilient, stress-tolerant, and nutritionally superior crop varieties will be made easier by these integrated methodologies, which will also improve our knowledge of complex features.

Rapid, accurate, and extensive evaluation of plant growth, physiological reactions, and stress adaption will be made possible by the integration of GS with high-throughput phenotyping and digital breeding platforms. The selection of superior genotypes can be accelerated by combining genomic and phenotypic data to better predict complex variables like yield, biomass, drought tolerance, and nitrogen-use efficiency. Similarly, deeper understanding of gene expression, metabolic pathways, stress responses, and the control of complex traits will be possible by multi-omics-assisted genomic selection, which incorporates transcriptomics, metabolomics, and epigenomics. Predicting agronomic performance, nutritional quality, and environmental adaption may all be improved by this integration.

By modelling complicated trait structures and nonlinear genetic interactions utilizing genomic, phenotypic, environmental, soil, and climatic data, artificial intelligence and machine learning will improve genomic prediction even further. Additionally, large-scale monitoring of crop growth, biomass, canopy design, disease prevalence, and stress responses will be made easier by unmanned aerial vehicles (UAVs), satellites, and remote sensing technology. By incorporating these data into genomic prediction models, it will enhance the evaluation of genotype $\times$ environment ($G \times E$) interactions and facilitate the creation of cultivars that are climate-smart and location-specific.

Additionally, the identification and targeted change of genes linked to yield, disease resistance, stress tolerance, and nutritional quality will be made easier by integrating GS with CRISPR-based genome editing, base editing, and prime editing. This strategy could increase selection efficiency and lessen the

need to assess huge edited populations. Through early genetic prediction and quick generation advancement, the integration of GS with speed breeding and controlled-environment agriculture would also shorten breeding cycles, enhancing genetic gain and hastening the production of superior cultivars.

5. Conclusion:

Genomic Selection (GS) is one of the promising breeding strategies that will become important in the future because the costs of genotyping will decrease, and powerful software will be developed. GS can be combined with gene editing, CRISPR-Cas9, phenomics, and multi-trait and multi-environment analysis. The accuracy of GS depends on the training population, which requires optimization. Better prediction algorithms, databases, and statistical methods can improve GS performance. Moreover, GS has to address genetic diversity, inbreeding, dominance, epistasis, and genetic background issues. In general, GS will play a critical role in crop breeding in the future due to the ability to overcome the mentioned challenges. Genomic Selection (GS) is one of the promising breeding strategies that will become important in the future because the costs of genotyping will decrease, and powerful software will be developed. GS can be combined with gene editing, CRISPR-Cas9, phenomics, and multi-trait and multi-environment analysis. The accuracy of GS depends on the training population, which requires optimization. Better prediction algorithms, databases, and statistical methods can improve GS performance. Moreover, GS has to address genetic diversity, inbreeding, dominance, epistasis, and genetic background issues. In general, GS will play a critical role in crop breeding in the future due to the ability to overcome the mentioned challenges.

References

1. Singh, B. D., Singh, A. K., & Singh, A. K. (2015). Marker-assisted plant breeding: principles and practices.
2. Bajad, S., Anil Kumar, A., Kumar, K., & Sharma, S. (2024). Genomic selection in plant breeding: A new era for crop improvement. *Agriculture Food: E-Newsletter*, 6(11), 579-582.
3. Sanu, K., Sikarwar, R. S., Sharma, S., Shukla, S., & Choubey, A. (2024). Genomic Selection for Multifaceted Characters in the Modern Era of Plant Breeding. In *Recent Advances in Plant Breeding-Volume 1* (pp. 150-165). Cornous Publications LLP.
4. Escamilla, D. M., Li, D., Negus, K. L., Kappelmann, K. L., Kusmec, A., Vanous, A. E., ... & Yu, J. (2025). Genomic selection: Essence, applications, and prospects. *The Plant Genome*, 18(2), e70053.