Plant Archives Vol. 25, No. 1, 2025 pp. 143-148 e-ISSN:2581-6063 (online), ISSN:0972-5210

Plant Archives

Journal homepage: http://www.plantarchives.org
DOI Url : https://doi.org/10.51470/PLANTARCHIVES.2025.v25.no.1.021

FUTURE DIRECTIONS IN WHEAT IMPROVEMENT: THE IMPACT OF GENOMIC SELECTION

P. N. Vinodh Kumar*, Sahana Police Patil and G. M. Keerthi

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi, India.
*Corresponding author e-mail: vinu49h20@gmail.com
(Date of Receiving-01-07-2024; Date of Acceptance-02-09-2024)

ABSTRACT

Wheat (*Triticum spp*.) is a crucial staple crop globally, contributing significantly to the caloric intake in many countries, including India. Traditional breeding methods have played a vital role in wheat improvement; however, they face limitations in addressing the complex challenges posed by biotic and abiotic stresses and the need for increased productivity. Genomic selection (GS) has emerged as a powerful tool to enhance wheat breeding by leveraging genomic information to predict breeding values with greater accuracy and efficiency. This review explores the application of GS in wheat improvement, highlighting its advantages, challenges, and future perspectives. Linear models like G-BLUP and Bayesian methods are extensively used in GS, offering simplicity and robustness. In contrast, non-linear models, including Random Forests, Support Vector Machines, and Artificial Neural Networks, capture complex genetic architectures, making them suitable for traits with intricate interactions. Despite its promise, GS faces challenges such as the need for large training populations, high costs, and integration with traditional breeding programs. Future directions for GS in wheat breeding include the development of climate-resilient varieties, the integration of artificial intelligence and machine learning to enhance prediction accuracy, and the combination of GS with precision agriculture for sustainable crop management. The continued advancement and adoption of GS in wheat breeding hold the potential to address global food security challenges by developing high-yielding and resilient wheat varieties.

***Key words*:** Genomic Selection (GS), Wheat Improvement, Breeding Values, Complex Traits, Climate Resilience

Introduction

Wheat (*Triticum spp*.) is an important staple cereal in many countries across the globe including India, which contributes approximately 20% of calories to the dietary requirement. Advances in plant breeding tools and techniques, mostly through conventional methods and agronomic approaches during the era of the green revolution and thereafter have contributed greatly to the annual productivity gain in wheat. The incorporation of dwarfing genes *Rht1* and *Rht2* in wheat cultivars by Borlaug and his team in the early 1960s was one of the most significant achievements to usher in the green revolution (Rajaram and Braun 2008). However, demand for wheat-based food products is increasing as the world's population grows, per capita income rises, and food consumption patterns become more diverse. Increasing the rate of genetic gain through modern breeding technologies is essential for food and nutritional security.

Wheat production faces significant challenges due to biotic and abiotic stresses, climate change, and the need for increased productivity to meet global demand. Traditional breeding methods have been instrumental in wheat improvement, but they are often time-consuming and limited by the complexity of traits such as yield and disease resistance. Genomic selection (GS) has emerged as a powerful tool that leverages genomic information to predict the performance of breeding lines, offering a more efficient and precise approach to wheat improvement (Heffner *et al*., 2009). Genomic selection (GS) is one such proven technology in animal breeding and has recently been incorporated into plant breeding programs, especially in the large-scale private sector. GS is a

promising approach for the rapid selection of superior genotypes and accelerating the breeding cycle. Traditional wheat breeding and yield improvement efforts are inadequate to cope with the 2% annual increment rate in the global population and feed an estimated ten billion population by 2050 (Hickey *et al.*, 2017). Traditional breeding methodologies rely on evaluating phenotypic merit along with pedigree information (Rasmusson and Phillips 1997) prompting lower accuracy and efficiency for trait selections that are modulated by prevailing environmental conditions (Heffner *et al.*, 2009) and hindering precision in selection. To overcome these challenges and to sustain production, modification and up gradation of conventional breeding techniques are prerequisites for meeting the production to feed the increasing population. Genomic selection is a form of MAS that simultaneously estimates all locus, haplotype, or marker effects across the entire genome to calculate genomic estimated breeding values (GEBVs; Meuwissen *et al.*, 2001). This approach contrasts greatly with traditional MAS because there is not a defined subset of significant markers used for selection. Instead, GS analyses jointly all markers in a population attempting to explain the total genetic variance with dense genome-wide marker coverage through summing marker effects to predict the breeding value of individuals (Meuwissen *et al.*, 2001).

With accurate genotypic and phenotypic information, genomic selection (GS) can facilitate the rapid selection and identification of desired genotypes by utilizing genome-wide distributed markers to estimate the effects of all loci and predict the genomic estimated breeding values (GEBV) to achieve more reliable selection. Linear models like G-BLUP and machine learning algorithms are used in understanding the complex patterns of data to make correct decisions. These prediction models can be effectively utilized in exploiting positive G×E interactions. Modelling multi-trait and multi-environment is a prerequisite for improving the prediction accuracy and performance of newly developed lines. The main advantages of GS over phenotype-based selection breeding are significant as it can facilitate accuracy in the selection, breeding time, and phenotyping costs in developing a variety, especially for complex traits with low heritability (Heffner *et al.*, 2009; Crossa *et al.*, 2017).

GS schemes are being implemented to attain genetic gains of economically important and low heritable traits which are otherwise very difficult to improve genetically by using conventional breeding principles. Incorporation and effective use of GS in the breeding program depends upon several factors such as breeding method, genetic architecture and heritability number of targeted traits, statistical models, availability of genotyping and phenotyping facilities, and the budget of the breeding program (Heffner *et al.*, 2009). Effective GS strategy utilizes an extensively genotyped and phenotyped population called a training population, which is used to optimize the statistical prediction model, with the help of which breeding values of the un-phenotyped population called as a breeding population are calculated called as genomic estimated breeding value (GEBV) purely based on genotyping data, which results in cutting down the breeding cycle and eliminating unnecessary multi-location and multi-environmental phenotyping trials.

The accuracy of these models depends on several factors, including the size and diversity of the training population, the density of markers used, and the genetic architecture of the trait being predicted. For complex traits like yield, which are influenced by many small-effect genes, dense marker coverage is essential to capture the genetic variance accurately. Recent advancements in high-throughput genotyping technologies, such as genotyping-by-sequencing (GBS) and SNP arrays, have made it feasible to generate large amounts of genotypic data, thus improving the robustness of genomic prediction models (Wang *et al.*, 2021).

Pipeline for genomic selection for wheat improvement

The implementation of GS in wheat breeding programs involves several critical steps. The selection of an appropriate training population is important for the success of genomic selection. The training population should be genetically diverse and representative of the breeding population to which the genomic prediction model will be applied. A well-chosen training population ensures that the model can capture the genetic variation present in the target population, leading to more accurate predictions.

In wheat, first step in genomic selection is to develop a training population, which is a diverse set of individuals for which both phenotypic data (traits of interest) and genotypic data (genetic markers) are available (Fig. 1).

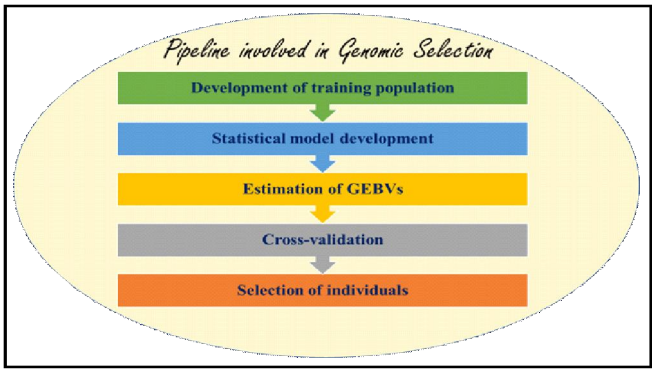

Fig. 1: Pipeline involved in the genomic selection.

The training population must be representative of the breeding population to ensure that the genomic prediction models are applicable (Jannink, *et al*., 2010). The size and composition of the training population are crucial, as they influence the accuracy of the prediction models. Ideally, the training population should cover a wide range of genetic diversity to capture the different genetic backgrounds present in the breeding population (Crossa *et al*., 2017).The next step involves developing a statistical model to predict the genetic potential of individuals based on their marker data. Various models can be used for genomic prediction.

Genomic selection (GS) models can be broadly classified into two main categories which are linear models and non-linear models. Linear models assume that the relationship between the markers and the trait of interest is linear. These models are often preferred for their simplicity, interpretability, and efficiency in handling large datasets. One among these are Best Linear Unbiased Prediction (BLUP) and its genomic variant (G-BLUP) are standard models in GS. G-BLUP uses genome-wide markers to estimate breeding values, if all markers contribute equally to genetic variance. The model assumes a normal distribution of marker effects, making it suitable for traits with many small-effect loci (VanRaden *et al*., 2008). The advantages of this model are Simple and robust, with balanced computational efficiency and widely used in both animal and plant breeding due to its reliability. The major limitation is it assumes equal contribution of all markers, which may not be suitable for traits controlled by major genes (Meuwissen *et al*., 2001). Various types of statistical model used in Genomic selection illustrated in Fig. 2.

Additionally, the Bayesian Ridge Regression (BRR) is a linear model that applies a Bayesian approach to the estimation of marker effects. Unlike G-BLUP, BRR allows for different variances for different markers. A model assigns normal priors to marker effects, which are then used to estimate the genomic breeding values.

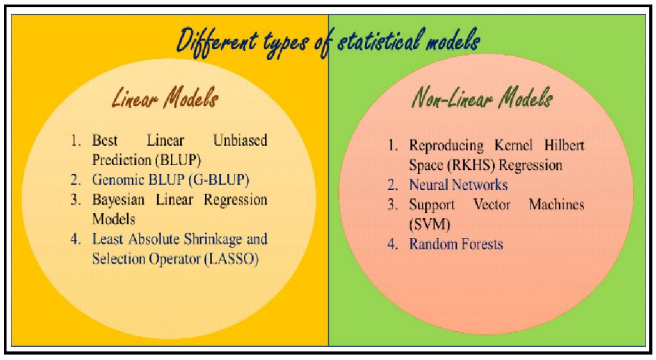

Fig. 2: various types of statistical model used in Genomic selection.

$$y=X\beta+Zg+e$$

Where, y is the vector of phenotypic values, X is the matrix of fixed effects, â is the vector of fixed effects, Z is the matrix of marker genotypes, g is the vector of random genetic effects, and õ is the vector of random residual effects (Meuwissen *et al*., 2001). The major advantages were more flexible than G-BLUP, allowing for marker-specific variances and Suitable for traits with varying effect sizes among markers. The key limitation was computationally more demanding than G-BLUP and it requires careful tuning of hyperparameters (Habierde *et al*., 2011)

Bayesian LASSO is a linear model that applies a Laplace prior to the marker effects. This prior induces a penalty that shrinks the effects of markers with small contributions, effectively performing variable selection. It is particularly useful for traits controlled by a few large-effect loci. The advantages of Bayesian LASSO are automatic variable selection reduces noise from markers with negligible effects and well-suited for traits with sparse genetic architecture. The limitations consist of computationally intensive, especially for large datasets and the choice of the shrinkage parameterë significantly influences the results (Gianola *et al*., 2013).

Bayesian models incorporate prior knowledge and provide a probabilistic framework for estimating marker effects. Common Bayesian methods include BayesA, BayesB, and BayesC, each differing in their assumptions about marker effect distributions (Gianola *et al*., 2008). These were encompassing a family of GS models that differ in their assumptions about the distribution of marker effects. Bayes A assumes different variances for each marker, Bayes B assumes a subset of markers have non-zero effects, and Bayes Cπ incorporates a probability that a marker has a non-zero effect. These models are particularly useful for traits with a mix of large and small-effect loci and the advantages were flexible and can accommodate different genetic architectures. It allows the incorporation of prior biological knowledge into the model. The limitations where it is computationally intensive, especially for large datasets. The choice of prior distributions and hyperparameters can be complex (Meuwissen *et al*., 2001; Habier *et al*., 2011).

Non-linear models capture complex interactions and non-linear relationships between markers and traits. These models are particularly useful for traits with intricate genetic architectures. It includes Random Forest (RF) model it is a machine learning model that creates an ensemble of decision trees. Each tree is built on a random subset of markers, and the final prediction is made by averaging the predictions of all trees. RF can capture

complex interactions between markers, making it suitable for traits with non-linear genetic architecture.It has capable of handling non-linear relationships and interactions between markers and often provides higher prediction accuracy for complex traits these were the advantages of RF. The major limitations are it requires large training datasets to achieve high accuracy and interpretation of the results can be challenging compared to linear models (Heslot *et al*., 2012).

Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks. In the context of GS, SVMs can handle high-dimensional genomic data by finding the optimal hyperplane that separates data points (genotypes) based on their associated phenotypic values. SVMs utilize kernel functions to transform input data into higher-dimensional spaces where linear separation is possible, effectively capturing non-linear relationships between markers and traits. It has some advantages like which is effective with high-dimensional data, suitable for datasets where the number of markers exceeds the number of samples (Long *et al*., 2011). It has good flexibility and robustness. Similarly, it has disadvantages also like computationally Intensive and training can be slow with large datasets. Performance is sensitive to the choice of kernel and hyperparameters and it is requiring extensive cross-validation. The obtained results can be difficult to interpret biologically.

Artificial Neural Networks (ANN) are computational models inspired by the human brain's network of neurons. ANNs consist of interconnected nodes (neurons) organized in layers that process input data to predict outputs. In GS, ANNs can model complex, non-linear relationships between genomic markers and phenotypic traits by learning from large datasets through training processes. The advantages are non-linear models are highly capable of capturing complex and non-linear interactions among genetic markers, making them particularly suitable for traits governed by intricate genetic architectures (González-Camacho *et al*., 2012). The adaptability of these models allows the network architecture, including depth and width, to be adjusted according to the complexity of the data, enabling better modeling of diverse traits. Additionally, these models often achieve high predictive accuracy, particularly when applied to large and diverse datasets. non-linear models have significant limitations, including the need for large training datasets to prevent overfitting and ensure the model's generalizability. The computational cost of training deep networks is another concern, as it requires substantial computational resources.

Table 1: Five-fold cross validation to evaluate the model.

Five-fold cross validation	**Subset 1**	**Subset 2**	**Subset 3**	**Subset 4**	**Subset 5**
Fold 1	Training set	Training set	Training set	Training set	Validation set
Fold 2	Training set	Training set	Training set	Validation set	Validation set
Fold 3	Training set	Training set	Validation set	Training set	Validation set
Fold 4	Training set	Validation set	Training set	Training set	Validation set
Fold 5	Validation set	Training set	Training set	Training set	Validation set

Genomic Estimated Breeding Values (GEBVs) are calculated for each individual in the breeding population using the developed statistical models. GEBVs represent the sum of the estimated effects of all markers across the genome, providing a prediction of an individual's genetic potential (VanRaden *et al*., 2008).

Cross-validation is a crucial step to evaluate the accuracy and reliability of the genomic prediction model. In cross-validation, the training population is divided into subsets, and the model is trained on a portion of the data (training set) and tested on the remaining data (validation set). This process is repeated several times, and the prediction accuracy is assessed by correlating the predicted GEBVs with the observed phenotypic values. The most common approach is k-fold cross-validation, where the data is split into k subsets, and the model is trained and tested k times, each time using a different subset as the validation set (Wimmer *et al*., 2013). Five-fold cross validation to evaluate the model tabulated in Table 1. Finally, individuals with the highest GEBVs are selected for further breeding. The goal is to choose individuals that carry the best genetic potential for the traits of interest, ensuring that these favourable traits are passed on to the next generation (Heffner *et al*., 2009).

Challenges and Limitations in Genomic Selection for Wheat Improvement

One of the primary challenges in genomic selection is the genetic diversity and population structure of the training populations. The accuracy of genomic selection models largely depends on the relatedness between the training population and the selection candidates. In wheat, a crop with a complex genome and significant genetic diversity across different varieties and breeding lines, population structure can introduce biases into the model. This bias may result in overestimation or underestimation of genetic values, ultimately affecting the accuracy of genomic estimated breeding values (GEBVs). Ensuring that the training population is representative of the breeding population is essential yet challenging due to the extensive genetic variation present in wheat. This challenge is

further compounded when breeding programs aim to combine diverse traits, such as stress tolerance and yield, which may have different genetic bases. Strategies such as using multi-environment trials and combining datasets across populations can help mitigate these issues, but they are resource-intensive and require careful consideration of population structure during model development (Crossa *et al*., 2017; Habier *et al*., 2007).

The implementation of genomic selection in wheat breeding programs demands substantial financial and logistical resources. High-throughput genotyping, which is essential for accurate marker-assisted selection, remains costly, particularly for large populations. Additionally, the development of reliable training populations and the subsequent phenotyping efforts to build predictive models require significant investments in time and resources. These requirements can be a barrier for smaller breeding programs or those in developing regions where funding and infrastructure may be limited. The costs associated with genomic selection can also limit the number of traits that can be selected simultaneously, as resources must be allocated to genotyping, model development, and validation for each trait. While advances in genotyping technologies and data analysis methods continue to reduce costs, the financial and resource barriers remain a significant limitation for the widespread adoption of genomic selection in wheat breeding (Hickey *et al*., 2014; Heffner *et al*., 2009).

Integrating genomic selection into existing breeding programs poses several challenges, particularly in terms of aligning new methods with traditional selection approaches. Breeders must balance the use of genomic selection with conventional phenotypic selection methods, especially in the early stages of breeding where phenotypic data is sparse or unreliable. Moreover, there is a need to adapt breeding pipelines to incorporate genomic data effectively, which may involve retraining personnel, updating software and computational infrastructure, and redesigning selection strategies. The transition to genomic selection also requires a cultural shift within breeding programs, as breeders must become comfortable relying on genomic data rather than traditional phenotypic evaluations alone. This integration is further complicated by the need to ensure that genomic selection models remain robust and accurate across different environments and over successive breeding cycles, which requires ongoing evaluation and adjustment (Bernardo, 2008; Jannink *et al*., 2010).

Future Perspectives and conclusions in Genomic Selection for Wheat Improvement

Genomic selection (GS) holds immense potential for accelerating the development of climate-resilient wheat varieties by enabling the selection of traits associated with stress tolerance at a genomic level. By incorporating genomic data related to these traits, breeders can more accurately predict which lines will perform well under adverse conditions, thus speeding up the breeding process. For instance, recent studies have shown that GS can effectively select for traits such as drought tolerance by utilizing genome-wide marker information to identify resilient genotypes (Juliana *et al*., 2019; Sehgal *et al*., 2020). This approach not only reduces the time needed to develop new varieties but also increases the likelihood of success in breeding for complex traits influenced by multiple genes and environmental factors. The integration of artificial intelligence (AI) and machine learning (ML) techniques into GS is opening new avenues for more accurate and efficient breeding programs. AI and ML can be used to enhance the predictive power of GS models by analysing large datasets, identifying complex patterns, and optimizing selection strategies. These technologies can also help in the identification of genomic regions associated with specific traits, making it possible to target breeding efforts more precisely. Moreover, AI-driven models can continuously learn and improve as more data becomes available, leading to increasingly accurate predictions of breeding values over time (Spindel *et al*., 2016). The use of AI and ML in GS is still in its early stages, but it has the potential to revolutionize wheat breeding by enabling more rapid and precise selection decisions.

The combination of GS with precision agriculture techniques offers a powerful approach to optimize wheat production in a sustainable manner. Precision agriculture involves the use of technologies such as remote sensing, GPS, and data analytics to monitor and manage crop growth at a fine scale. When integrated with GS, precision agriculture can provide real-time phenotypic data that enhances the accuracy of genomic predictions. This integration allows breeders to account for environmental variability more effectively and to select varieties that are well-suited to specific growing conditions. For example, precision agriculture can help identify micro-environments within fields that are particularly challenging for crop growth, and GS can be used to develop varieties tailored to those conditions (Moser *et al*., 2019; Araus *et al*., 2018). This synergy between GS and precision agriculture could lead to more efficient use of resources, reduced environmental impact, and higher yields.

Conclusion

Genomic selection is transforming wheat breeding by enabling the selection of complex traits with greater

accuracy and efficiency. However, challenges such as genetic diversity, cost, and integration with existing breeding programs must be addressed to fully realize its potential. Future perspectives indicate a promising role for GS in developing climate-resilient wheat varieties, with AI and ML enhancing model precision, and precision agriculture providing valuable phenotypic data for improved selection accuracy. The future of wheat improvement through genomic selection looks promising, with the potential to meet global food security challenges in the face of climate change. By leveraging advances in genomics, AI, and precision agriculture, breeders can develop wheat varieties that are not only higher yielding but also more resilient to environmental stresses. Continued research and investment in these areas will be crucial to unlocking the full potential of genomic selection in wheat improvement, ensuring that future generations have access to sustainable and nutritious food sources.

References

Araus, J.L., Cairns J.E. and Simón M.R. (2018). Genomic prediction of optimal wheat ideotypes for challenging environments. *Global Food Security*, **19**, 10-18.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, **48(5)**, 1649-1664.

Crossa, J., *et al.* (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, **22(11)**, 961-975.

Crossa, J., Pérez P., Hickey J., Burgueno J., Ornella L., Cerón-Rojas J. and Campos G.D.L. (2017). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, **112(1)**, 48-60.

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, **194(3)**, 573-596.

Gianola, D., de los Campos G., Hill W.G., Manfredi E. and Fernando R. (2008). Additive genetic variability and the Bayesian alphabet. *Genetics*, **179(3)**, 1747-1768.

González-Camacho J.M., Crossa J., Pérez-Rodríguez P., Ornella L., Gianola D. and Atlin G. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, **125(4)**, 759-771.

Habier, D., Fernando R.L. and Dekkers J.C.M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, **177(4)**, 2389-2397.

Heffner, E.L., Sorrells M.E. and Jannink J.L. (2009). Genomic selection for crop improvement. *Crop Science*, **49(1)**, 1-12.

Heslot, N., Yang H.P., Sorrells M.E. and Jannink J.L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Science*, **52(1)**, 146-160.

Hickey, J.M., Chiurugwi T., Mackay I., Powell W. and Hayes B.J. (2014). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, **46(2)**, 124-130.

Hickey, J.M., Chiurugwi T., Mackay I. and Powell W. (2017). Implementing Genomic Selection in CGIAR Wheat Breeding Programs Workshop Participants. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics*, **49(9)**, 1297-1303.

Jannink, J.L., Lorenz A.J. and Iwata H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, **9(2)**, 166-177.

Juliana, P., Singh R.P., Poland J., Mondal S., Crossa J., Montesinos-Lopez O. and Guzmán C. (2019). Prospects and challenges of applied genomic selection—A new paradigm in breeding for climate resilience in wheat. *Frontiers in Plant Science*, 1**0**, 617.

Long, N., Gianola D., Rosa G.J., Weigel K.A. and Avendaño S. (2011). Machine learning classification procedure for selecting SNPs in genomic selection: Application to early mortality in broilers. *Journal of Animal Breeding and Genetics*, **128(2)**, 109-119.

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(4)**, 1819-1829.

Moser, G., Tier B., Crump R.E., Khatkar M.S. and Raadsma H.W. (2019). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, **41(1)**, 1-13.

Rajaram, S. and Braun H.J. (2008). Wheat yield potential. In International Symposium on Wheat Yield Potential: Challenges to International Wheat Breeding (103-107). CIMMYT (International Maize and Wheat Improvement Center), Mexico, Mexico.

Sehgal, D., Mondal S., Crespo-Herrera L., Velu G., Juliana P., Huerta-Espino J. and Singh R.P. (2020). Fifty years of semi-dwarf spring wheat breeding at CIMMYT: Grain yield progress in optimum, drought, and heat stress environments. *Field Crops Research*, **250**, 107757.

Spindel, J., Begum H., Akdemir D., Collard B., Redona E., Jannink J.L. and McCouch S.R. (2016). Genomic selection and association mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLOS Genetics*, **11(4)**, e1004982.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91(11)**, 4414-4423.

Wang, X., Xu Y., Hu Z. and Xu C. (2021). Genomic selection methods and their applications in plant breeding. *Current Genomics*, **22(3)**, 172-183.

Wimmer, V., Albrecht T., Auinger H.J. and Schön C.C. (2013). Synbreed: A framework for the analysis of genomic prediction data using R. *Bioinformatics*, **28(15)**, 2086-2087.