Eco. Env. & Cons. 28 (1) : 2022; pp. (512-517)
Copyright@ EM International
ISSN 0971–765X

DOI No.: http://doi.org/10.53550/EEC.2022.v28i01.077

Effects of marker density and training set size on the genomic selection accuracy for predicting sheath blight resistance in rice (*Oryza sativa* L.)

Mahantesh[1*], K. Ganesamurthy[2], Sayan Das[3], R. Saraswathi[4], C. Gopalakrishnan[5] and R. Gnanam[6]

[1]Centre for Plant Breeding and Genetics, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[2, 4]Department of Rice, Centre for Plant Breeding and Genetics, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[3] Molecular Breeding Lead, Pioneer Hi-Bred Private Limited, Tunkikalsa Village, Medak 502 336, Telangana, India
[5]Department of Plant Pathology, Centre for Plant Protection Studies, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India
[6]Department of Plant Molecular Biology & Bioinformatics, CPMB, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

(Received 5 December, 2021; Accepted 17 January, 2022)

ABSTRACT

The success of genomic selection mainly depends on the extent of linkage disequilibrium between markers and quantitative trait loci, size of training set, heritability of the trait etc. The extent of linkage disequilibrium depends on the genetic structure of the population and marker density. This study was conducted to determine the effects of marker density and size of training set on prediction accuracy using 1545 recombinant inbred lines derived from eleven bi-parental rice populations. All RILs were genotyped with 6564 SNPs and screened in two hot spot locations to assess reaction against sheath blight. Bayesian B model was used to train the statistical model for calculation of marker effects and genomic estimated breeding values. To evaluate the genomic prediction accuracy, various levels of training set size (300, 500, 700, 900 and 1200 lines) and marker density (500, 800, 1100, 1400, 1700, 2000, 4000 and 6000 markers) were considered. In our study, the prediction accuracy increased with increase in training set size, however, average prediction accuracy of 0.717 was obtained for the training set comprising of 900 lines before reaching plateau with marginal increase in prediction accuracy with higher training set sizes. The predictive ability increased dramatically with more SNPs included in the analysis until 2000 markers with average prediction accuracy of 0.681, no significant improvement beyond this was observed. The results indicate that training set with approximately 900 lines and 2000 uniformly distributed SNP markers with good amount of polymorphism across populations would be enough to reach achievable accuracy to predict sheath blight resistance in rice.

***Key words**: Rice, SNP, Genomic selection, Sheath blight, Bayesian B*

([1*]Ph.D Student, [2] Professor and Head, [4] Professor (PB&G)

Introduction

Sheath blight is considered as one of the devastating diseases of rice worldwide leading to significant yield losses in many rice growing counties, it is caused by a necrotrophic pathogen Rhizoctoniasolani (Rao *et al*., 2020). Because of unique symptoms exhibited by this disease it is referred as "rotten foot stalk", "mosaic foot stalk" and "snake skin disease" (Molla *et al*., 2020; Zhang *et al*., 2019b).

The most economic and effective strategy in order to control the disease is, development of cultivars with resistance to sheath blight but only few varieties are resistant and few reliable QTLs have been discovered so far which are linked to sheath blight resistance (Chen *et al.,* 2019). Because of lack of good number of authentic and reliable sources of resistance, breeding for sheath blight has been challenging in Rice (Zuo *et al*., 2010; Srinivasachary, *et al.,* 2011). Upon intensive study sheath blight is believed to be controlled by many quantitative trait loci scattered across the genome (Zuo *et al.,* 2013). It is widely believed that quantitative nature of resistance could be the expedient for evolving varieties with durable/horizontal resistance (Heslot *et al*., 2013).

When the target trait is complex, limited genetic gain may be expected because genetics is controlled by many genes with smaller effects, which means many number of markers explain the genetic variance (Bernardo, 2008). In 2001, Meuwissen *et al*. proposed the conceptof Genomic Selection (GS), which assumes that many genomic regions contribute to the genetic variance and each region is in LD (linkage disequilibrium) with at least one known marker. If SNPs markers are used, effects of all SNPs dispersed across the genome are estimated and used for predicting genetic value of the selection candidates. The GS has become more popular because of advances in the genotyping technology which led to reduction in the cost involved in high throughput genotyping, thus remarkable improvement has been observed in terms of selection accuracy and eventually helped in reducing breeding cycle and increase in genetic gain. (Bhat *et al*., 2016; Meuwissen *et al*., 2016; Crossa *et al*., 2017; Weller *et al*., 2017).

Besides the statistical models used for training the dataset, the prediction accuracy may also be impacted by marker density, genetic architecture of the target trait, minor allele frequency, heritability, training set size, LD between markers and QTLs etc. Also the prediction accuracy can be increased by usage of high density markers which are uniformly distributed and cost effective genotyping technology adds to it (Elshire *et al*., 2011). The current investigation was carried out to understand the effect of marker density and size of the training set on prediction accuracy so that a balanced strategy can be followed for predicting sheath blight resistance without making tradeoff among computational efficiency, cost involved in phenotyping and prediction accuracy for successful deployment of genomic selection in practical plant breeding programs.

Materials and Methods

Parent material and phenotyping of F_7 RILs for ShB

The material used for the existing study comprised of 1545 RILs from eleven bi-parental populations formed by crossingresistant lines with agronomically superior susceptible linesinvolving Jasmine 85, Tetep & MTU 9992 as resistant parents and TN1, Swarna-Sub1, II32B, IR54 & IRBB4 as susceptible parents. The RILs were created by following single seed descent method (SSD) at Rapid Generation Advancement/ Speed breeding facility of Pioneer Hi-Bred Pvt. Ltd. Research Centre at Tunkikalsa village, Medak district, Telangana. The eleven crosses utilized for the study were, Jasmine 85×TN1, Jasmine 85×Swarna-Sub1, Jasmine 85×II32B, Jasmine 85×IR54, Tetep×TN1, Tetep×Swarna-Sub1, Tetep×II32B, Tetep×IR54, MTU 9992×TN1, MTU 9992×II32B and MTU 9992×IRBB4. All the RILs were phenotyped for sheath bight reaction in two hot spot locations (Seethanagaram and Draksharam) of East Godavari District of Andhra Pradesh state, India (Latitude $16^0$08′ N and Longitude $81^0$08′ E, Latitude $17^0$10′N and Longitude $81^0$41′ E).

The experiments containing F7 progenies along with parental lines were planted in Randomized complete design with two replications. Row length of 1.2 meter and spacing of 15 cm ×10 cm was considered to ensure dense population which is amiable for the development of disease. TN1 was used as susceptible check and was sown after every two rows as well as all along the border to upsurge the disease pressure as to serve as spreader rows. In the current study, the virulent local East Godavari isolate of rice sheath blight pathogen was utilized for disease screening. Before the inoculation, the fungus

was cultivated in potato dextrose agar medium at optimal temperature for 3–4 days, followed by transferring of disc of medium with mycelia for increase. To ensure rigorous screening for better disease development, artificial inoculation was done by spraying the mycelia uniformly at the base of plant at maximum tillering stage. The data was recorded at peak milking stage to dough stage by visualizing the relative lesion length to height (%) using 1–9 scale (SHBSC - Sheath blight score) based on development of lesion from the lower to upper part of plant on a scale from 1 (Resistant) to 9 (Susceptible) thereby getting total of six phenotypic categories, where score 1: no infection, score 2:1–20%, score 3: 21–30%, score 5: 31–45%, score 7: 46–65%, score 9: 66–100%.

SNP genotyping

All the RILs used for the study were genotyped using Infinium-XT marker platform which is a fixed plex comprising of 6564 markers, the genotyping was done at marker technology lab of Pioneer Hi-Bred International Limited at Johnston, Iowa State, United States of America.

GS modeling

Genomic selection follows a three-step process. First, all the individuals which are part of training set are genotyped and phenotyped and effects are estimated for all molecular markers, GEBVs (predicted values) were calculated for all the individuals which are part of same training set using the marker effects generated and were correlated with phenotypic values to get prediction accuracy, this is referred as data fit analysis of training set. Second, the training set is validated by considering independent data set, different approaches of cross validation are used to understand predictive ability of training set. Third, members of untested populations are solely genotyped and then selected based on their predicted phenotypes (GEBVs) according to the marker effects estimated in the training set.

For the current investigation, Bayesian B model was used for statistical analysis to generate marker effects to get GEBV's of the breeding lines.The statistical analysis was done in "R" program with BGLR package with 50,000 iterations.

Cross-validation method to study impact of marker density(MD) on prediction accuracy

To evaluate the effect of marker density (MD) on the accuracy of prediction, various levels of marker density were considered (500, 800, 1100, 1400, 1700, 2000, 4000 and 6000 markers) and in order to assess the ability of genomic prediction for eachSNP set with different density, ten-fold cross-validation method was utilized. Wherein,1545lines were randomly and evenly divided into 10 subsets. Oneof the 10 subsets was used as the validation set and theremaining 9 subsets were used as the training set (training and validation set comprise of 1390 and 155 lines respectively). GEBVs for the lines present in the validation set were estimated using the marker effects generated from training set with Bayes B method. The predictive ability was assessed by calculating the correlation between GEBVs and phenotypic values for the lines present in validation set.The procedure was repeated ten times ensuring that each subset was used as validation set at least once, finally prediction accuracy (correlation coefficients) values across ten-fold were averaged. The ten-fold cross validation was repeated for all the eight SNP setsconsidered with varying marker densities.

Cross-validation method to study impact of training set size (TSS) on prediction accuracy

To evaluate the effect of training set size (TSS) on the accuracy of prediction, various levels of training set sizes were considered (300, 500, 700, 900 and 1200 lines) with constant marker density (6564 markers) and in order to assess the ability of genomic prediction for each training set with different size of lines, ten-fold cross-validation method was utilized. Where lines present in each set were randomly and evenly divided into 10 subsets. One of the 10 subsets was used as the validation set and the remaining 9 subsets were used as the training set. GEBVs for the lines present in the validation set were estimated using the marker effects generated in training set with Bayes B model. The predictive ability was assessed by calculating the correlation between GEBVs and phenotypic values for the lines present in validation set. The procedure was repeated ten times ensuring that each subset was used as validation set at least once, finally prediction accuracy (correlation coefficients) values across ten-fold were averaged. The ten-fold cross validation was repeated for all the five training sets considered with varying sizes.

Results and Discussion

The frequency distribution of 1545 RILs evaluated

showed continuous variation across all population studied for sheath blight (Figure 1). The genotypic analysis was done with large number of markers which were uniformly distributed throughout the genome (Table 1), polymorphic markers between parents across populations studied ranged from 1407 to 2849, MTU 9992×TN1 and MTU 9992×IRBB4 possessed lowest and highest number of informative markers (Table 2).

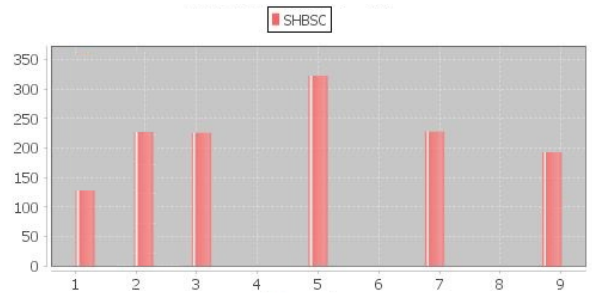

Fig. 1. Frequency distribution of sheath blight phenotypic scores

Effect of marker density (MD) on prediction accuracy estimation

The effect of marker density on prediction accuracy was assessed through random ten-fold cross validation with Bayes B model. The analysis was done keeping the training and validation set size constant (1390 and 155 lines respectively). The average prediction accuracy across ten-fold cross validation obtained was 0.336 with MD = 500, 0.443 with MD = 800, 0.471 with MD = 1100, 0.535 with MD = 1400, 0.625 with MD = 1700, 0.681 with MD = 2000, 0.698 with MD = 4000, and 0.708 with MD = 6000.Prediction accuracy improved as the MD increased, a strong response to increase in marker density up to 2000 markers was observed with only a marginal increase in prediction accuracy when increased from 2000 to 6000 markers. The results are summarized in Figure 2. The box plot clearly disclosed that there was high range of prediction accuracy values across tenfold especially for lower MD datasets and values were not consistent (Figure 3). Results indicated that 2000 markers were enough for generating a relatively accurate prediction calibration within the panel of lines used for the study.The results were consistent with studies using smaller data sets where additional markers benefited in enhancing prediction accuracy when larger training sets were used (Heffner *et al*., 2011a,b).

Table 1. Summary of marker data used for analysis and SNPs distribution on each chromosome

Chromosome	SNPs	Length (cM)
Ch1	639	181.8
Ch2	846	162.84
Ch3	598	164.04
Ch4	594	129.6
Ch5	583	128.58
Ch6	577	124.4
Ch7	457	118.6
Ch8	495	121.2
Ch9	427	93.1
Ch10	324	84.01
Ch11	541	117.9
Ch12	483	109.5
Total	**6564**	**1535.47**

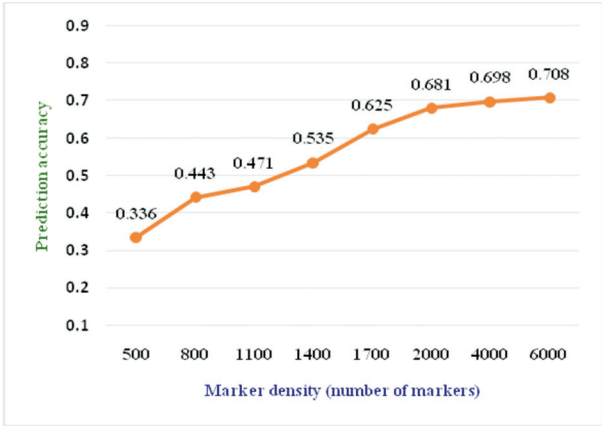

Fig. 2. Line graph depictingaverage prediction accuracies of different levels of marker densities studied with Bayes B model

Effect of training set size (TSS) on prediction accuracy estimation

The effect of training set size on prediction accuracy was evaluated through random ten-fold cross validation with Bayes B model. The analysis was performedkeeping the marker density (MD) constant (6564 markers) and all the SNPs were used for prediction. The average prediction accuracyacross ten-fold cross validation obtained was 0.468 with TSS = 300, 0.577 with TSS = 500, 0.678 with TSS = 700, 0.717 with TSS = 900, and 0.722 with TSS = 1200. The results are summarized in Figure 4.The box plot clearly revealed that there was range of prediction accuracy values across tenfold for each TSS and values were not consistent (Figure 5).Prediction accuracy increased as the TSS increased, a sharp in-

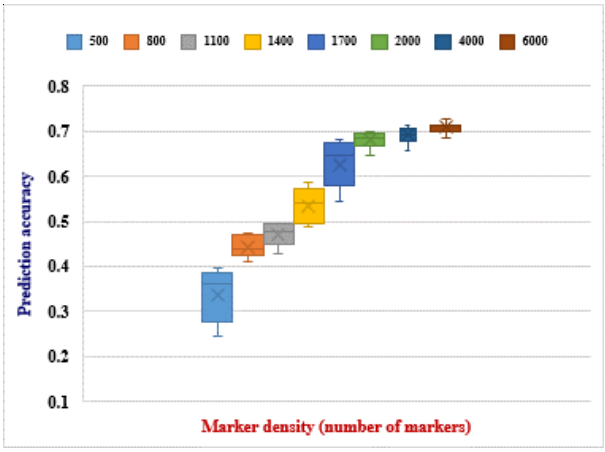

Fig. 3. Box plot depicting ten-fold cross validation results of different levels of marker densitiesstudied with Bayes B model

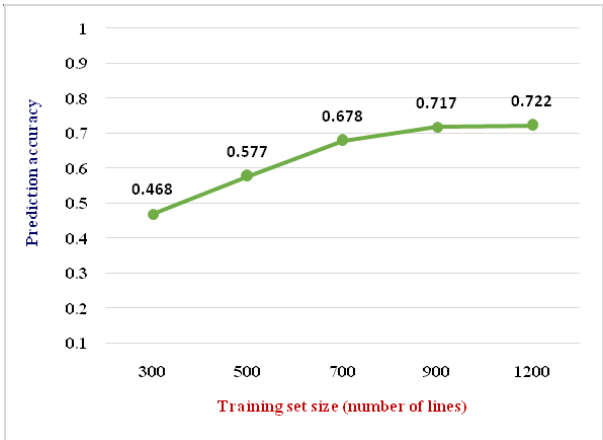

Fig. 4. Line graph depictingaverage prediction accuracies of different levels of training set sizes studied with Bayes B model

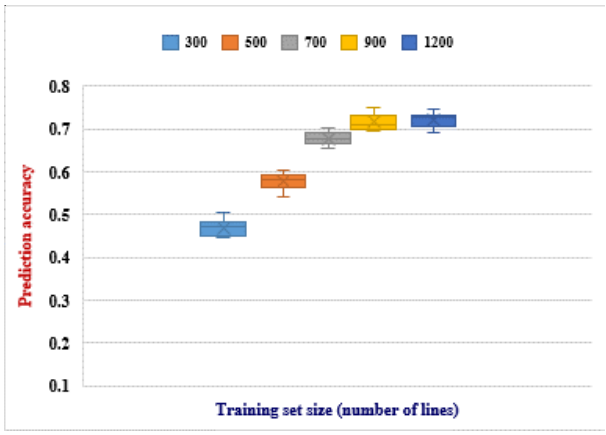

Fig. 5. Box plot depicting ten-fold cross validation results of different levels of training set sizes studied with Bayes B model

crease in accuracy was found before reaching a plateau at TSS approximately 900 lines, with only a marginal increase in prediction accuracy when TSS increased from 900 to 1200 lines. It indicated that a training set with approximately 900 lines can provide the maximum achievable accuracy, hence resource allocated for phenotyping could be reduced with optimum size of training set. This result confirms previous findings from smaller populations (Heffner *et al*., 2011a, b; Isidro *et al*., 2015), and extends the relationship to larger training sets showing there is a point at which accuracy begins to plateau in response to increased training set size (Adam *et al*., 2018).

Conclusion

From the current investigation we could observe

Table 2. The informative markers available across the genome for each population used for analysis

Populations	Number of RILs	Total Markers	Polymorphic Markers
Jasmine 85/TN1	121	6564	2522
Jasmine 85/Swarna-Sub1	139	6564	2627
Jasmine 85/II32B	144	6564	2586
Jasmine 85/IR54	161	6564	2663
Tetep/TN1	221	6564	2806
Tetep/Swarna-Sub1	158	6564	2278
Tetep/II32B	241	6564	2702
Tetep/IR54	94	6564	2796
MTU 9992/TN1	50	6564	1407
MTU 9992/II32B	122	6564	2314
MTU 9992/IRBB4	94	6564	2849
Total	**1545**		

that there was a point at which prediction accuracy begins to plateau in response to training set size, and this response was independent from the genetic complexity of the trait. Also, we detected that the response to increased marker density was larger when using a diverse training set and predicting from poorly related training sets. This indicates that the high-density genotyping platforms are need of the hour for successful deployment of genomic selection for complex traits like sheath blight whose inheritance is governed by multiple genes. The investigation provides great idea for pragmatic plant breeders to optimally design their genomic selection strategy to achieve high selection accuracy and subsequent rates of genetic gain.

Acknowledgements

I would like to thank Corteva Agriscience (Pioneer Hi-Bred Private Limited, Tunkikalsa Village, Medak, Telangana State, India) for providing all the facilities to carry out my research work. I greatly acknowledge my advisory committee members for their suggestions, support and guidance.

References

Adam, N., Julian, T., James, E. and Haydn Kuchel. 2018. Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3: Genes, Genomes. Genetics.* 8(9): 2889–2899.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci.* 48 : 1649.

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S. and Jadon, V. 2016. Genomic selection in the Era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* 7: 221.

Chen, Z., Feng, Z., Kang, H., Zhao, J. and Chen, T. 2019. Identification of new *resistance loci against sheath blight disease in rice through genome-wide association study. Rice Science.* 26(1) : 21–31.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D. and de Los Campos, G. 2017. Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22 : 961–975.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K. and Buckler, E. S. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.

Heffner, E., Jannink, J., Iwata, H., Souza, E. and Sorrells, M. 2011a. Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51: 2597–2606.

Heffner, E., Jannink, J. and Sorrells, M. 2011b. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4: 65–75.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J. L. and Sorrells, M.E. 2013. Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. *PLoS One*. 8(9): e74612.

Isidro, J., Jannink, J., Akdemir, D., Poland, J. and Heslot, N. 2015. Training set optimization under population structure in genomic selection. *Theor. Appl. Genet*. 128: 145–158.

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

Meuwissen, T., Hayes, B. and Goddard, M. 2016. Genomic selection: a paradigm shift in animal breeding. *Anim. Front*. 6: 6–14.

Molla, K.A., Karmakar, S., Molla, J., Bajaj, P., Varshney, R. K., Datta, S. K. and Datta, K. 2020. Understanding sheath blight resistance in rice: the road behind and the road ahead. *Journal of Plant Biotechnology*. 18 : 895–915.

Rao, T. B., Chopperla, R., Prathi, N. B., Balakrishnan, M., Prakasam, V., Laha, G. S., Balachandran, S. M. and Mangrauthia, S. K. 2020. A comprehensive gene expression profile of pectin degradation enzymes reveals the molecular events during cell wall degradation and pathogenesis of rice sheath blight pathogen *Rhizoctoniasolani*. *Journal of Fungi* 6 : 71–82.

Srinivasachary, L., Willocquet, L. and Savary, S. 2011. Resistance to rice sheath blight (*Rhizoctoniasolani* Kuhn) [teleomorph: *Thanatephoruscucumeris* (A.B. Frank) Donk.] disease: Current status and perspectives. *Euphytica* 178 : 1-22

Weller, J. I., Ezra, E. and Ron, M. 2017. Invited review: a perspective on the future of genomic selection in dairy cattle. *J. Dairy Sci.* 100 : 8633–8644.

Zhang, S.W., Yang, Y. and Li, K.T. 2019b. Occurrence and control against rice sheath blight. *Biology of Disease Science* 42: 87–91.

Zuo, S. M., Zhang, Y. F., Chen, Z. X., Chen, X. J. and Pan, X. B. 2010. Current progress on genetics and breeding in resistance to rice sheath blight. *Scientia Sin Vitae.* 40 : 1014–1023.

Zuo, S. M., Yin, Y. J., Zhang, L., Zhang, Y.F., Chen, Z. X. and Pan, X. B. 2013. Fine mapping of q*SB-11*, the QTL that confers partial resistance on rice sheath blight. *Theoretical and Applied Genetics.* 126: 1257–1272.