Analysis of the Efficiency of Genomic Selection Models for Predicting Sheath Blight Resistance in Rice (*Oryza sativa* L.,)

Mahantesh[1]✉️📷, K. Ganesamurthy[2], Sayan Das[3], R. Saraswathi[2], C. Gopalakrishnan[3] and R. Gnanam[4]

[1]Centre for Plant Breeding and Genetics, [2]Dept. of Rice, Centre for Plant Breeding and Genetics, [3]Dept. of Plant Pathology, Centre for Plant Protection studies, [4]Dept. of Plant Molecular Biology & Bioinformatics, CPMB, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu (641 003), India
[3]Pioneer Hi-Bred Private Limited, Tunkikalsa Village, Medak, Telangana State (502 335), India

Open Access

Corresponding ✉️ mgpatil_3401agri@yahoo.com

📷 0000-0003-2544-0566

ABSTRACT

The research was undertaken during June-October 2020 at Seethanagaram and Draksharam villages of East Godavari district, Andhra Pradesh, India with an objective to evaluate efficiency of genomic selection models involving 1545 recombinant inbred lines (RILs) derived from eleven bi-parental populations in Rice. During June-October 2020, the F_7 RILs were screened in two hot spot locations. The genotyping was done with Infinium platform having 6564 SNP markers. Five models were used rrBLUP, BayesA, BayesB, BayesCPi and GBLUP to train the statistical model for calculation of marker effects and genomic estimated breeding values (GEBVs). The prediction accuracy (data fit) of training set across models ranged 0.63–0.72, lowest and highest prediction accuracies were observed with rrBLUP and GBLUP models respectively. Tenfold cross validation with different approaches, the average prediction accuracy ranged from 0.60 (rrBLUP, BayesA, BayesB and BayesCPi) –0.72 (GBLUP). BayesB and GBLUP models exhibit higher prediction accuracies compared to other models studied. The predictive ability increased dramatically with more SNPs included in analysis until 2000 markers with average prediction accuracy of 0.681, no significant improvement beyond this was observed. The results are lucrative, all in all, high prediction accuracies observed in this study suggest genomic selection as a very promising strategy while breeding for sheath blight resistance in rice to increase genetic gain.

KEYWORDS: Rice, sheath blight, SNP, genomic selection

1. 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 *Rhizoctonia solani* (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).

This disease has become popular recently because of intensification of rice-cropping systems with the usage of higher amount of nitrogen and development of rice cultivars with high yielding, high tillering and semi-dwarf stature which suit to high plant densities (Yellareddygari et al., 2014). In India its prevalence is mainly confined to coastal places where farmers grow high yielding varieties. Frequent rainfall, high temperature (28–32°C) coupled with high humidity 95–97% favors disease development hence the disease is very common in rainy season in India (Amandeep et al., 2015).

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, Willocquet and Savary, 2011). Upon intensive study it's 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 (Poland et al., 2013).

Because of complex inheritance it very difficult to exploit and tap the genomic regions using classical approaches of QTL mapping (linkage and LD analysis). One of popular approach which is very popular now a days which can help in breeding for complex traits is genomic selection (GS). Genomic selection uses large number of markers scattered across the genome which are in LD with many genomic regions of interest (Meuwissen et al., 2001). It has been shown to be effective for improving quantitative traits, both in simulations (Bernardo and Yu, 2007) and in empirical studies (Heslot et al., 2013; Lorenz et al., 2012; Rutkoski et al., 2011, 2012 and 2014). In many studies bi-parental populations with good genetic relationship between the training and test populations have exhibited better prediction accuracies in comparison with three way and complex cross populations, the reason could be controlled population structure and greater linkage disequilibrium between markers and QTLs (Bernardo and Yu, 2007).

The prediction accuracy of training set relies heavily on many factors like, genetic relationship between populations which are part of training set and test set, marker density (number and distribution of markers), size of the training set, statistical models used for analysis etc. The current investigation was done with recombinant inbred lines developed from eleven bi-parental populations to study the efficiency of different statistical models used for genomic selection and the effect of marker density on accuracy to predict sheath blight resistance in rice.

2. MATERIAL AND METHODS

2.1. Parent material and phenotyping of F_7 RILs for ShB

A total of 250 germplasm lines were screened for identification of lines which are resistant and susceptible to Sheath blight disease during June – October 2016 at Seethanagaram village of East Godavari district, Andhra Pradesh, India (Latitude $16^008'$ N and Longitude $81^008'$ E) by pathology team of Pioneer Hi-Bred Private Limited. Based on earlier studies and information available in public domain, lines were selected and crosses were made involving Jasmine 85, Tetep & MTU 9992 as resistant parents and TN1, Swarna Sub1, II32B, IR54 & IRBB4 as susceptible parents. The total of 1545 RILs from eleven bi-parental populations were used for the study to tap all the genomic regions governing sheath blight resistance dispersed across the genome. The RILs were generated 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 used 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 $16^008'$ N and Longitude $81^008'$ E, Latitude $17^010'$N and Longitude $81^041'$ E).

The experiments consisting of F_7 progenies along with parental lines were planted in Randomized complete design with two replications. Row length of 1.2 meter and spacing of 15×10 cm² was considered to ensure dense population which is congenial 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 increase the disease pressure as to serve as spreader rows. In the present 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 multiplication. To ensure stringent 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 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: 1–20%, score 3: 21–30%, score 5: 31–45%, score 7: 46–65%, score 9: 66–100%.

2.2. SNP genotyping

All the RILs used for the study were genotyped using Infinium 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.

2.3. Statistical analysis (GS modeling)

Genomic selection follows a three-step process (Figure 1). 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 the training set. Second, the training set is cross-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 rrBLUP, BayesA, BayesB, BayesCPi and GBLUP models were used for training the model and 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.

2.4. Tenfold cross validation analysis

To assess the accuracy of the model's ability to predict the untested lines tenfold cross-validation (CV) simulations were done. The training set comprising of 1545 lines from eleven population with phenotypic and genotypic data was used to perform repeated tenfold cross validation. The training set was randomly divided into ten portions, the statistical model was trained on nine portions with 1390 lines (training set), the remaining 155 lines (validation set) GEBVs were predicted using the marker effects of nine portions training set, these GEBV of 155 lines were correlated with phenotypic values to know the prediction accuracy. All these steps were repeated ten times to ensure that each portion was used at least once for prediction of GEBVs, finally the obtained accuracies across tenfold were averaged to understand the predictive ability of training set with all the models used for the study.

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). The ten-fold cross validation procedure was repeated for varying marker density datasets with GBLUP and Bayes B models.

3. RESULTS AND DISCUSSION

The frequency distribution of 1545 F7 RILs evaluated showed continuous variation across all population studied (Figure 2). 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). The number of markers (marker density) and distribution of markers were found to have great impact on the prediction accuracy of training set, more the markers
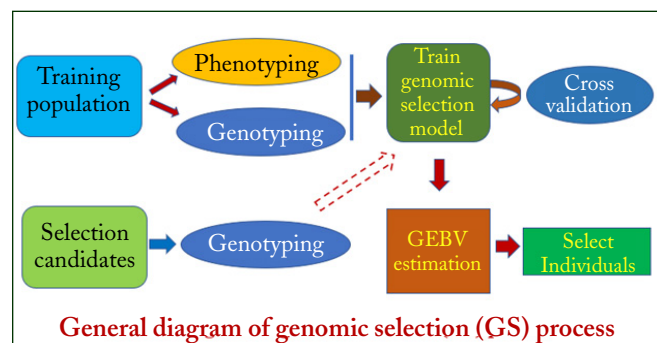

Figure 1: Showing the different steps of genomic selection (GS) used for crop improvement program
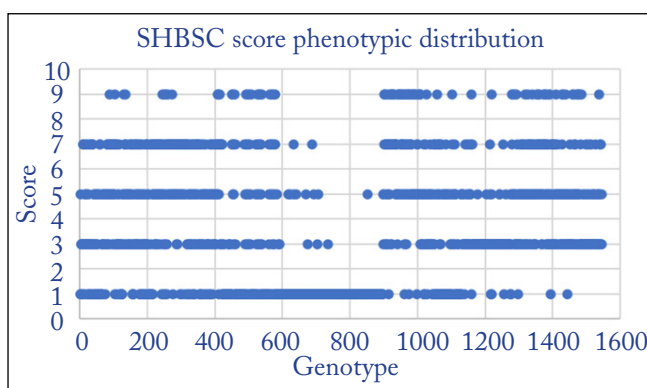

Figure 2: Distribution of sheath blight phenotypic scores into five classes or categories

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
Ch10	324	84.01
Ch11	541	117.9
Ch12	483	109.5
Total	6564	1535.47

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

Populations	No. 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		

in LD with QTLs governing the trait, higher prediction accuracy was obtained.

The prediction accuracy (data fit) of training set across five models studied (rrBLUP, BayesA, BayesB, BayesCPi and GBLUP) ranged 0.63–0.72, lowest and highest prediction accuracies were observed with rrBLUP and GBLUP models respectively (Figure 3). BayesB and GBLUP models exhibited similar prediction accuracies with no significant difference between them. Greater data fit results of the training set in the current study could be attributed to larger size of the training set with bi-parental populations
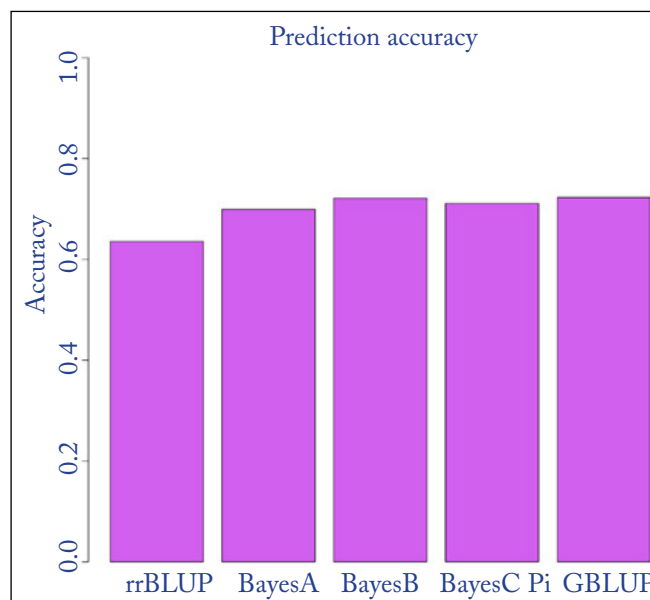

Figure 3: Prediction accuracy results (data fit) across populations for all the individuals which were part of training set for all five models studied

having good number of progenies in each population (Table 2), higher marker density, high LD between markers and QTLs, robust statistical models used for calculation of marker effects (except rrBLUP), greater genetic relationship between populations which were part of training and validation set. The results were in conformation with the results obtained in earlier studies on effect of marker density, size of training set etc. (Heffner et al., 2011a; Heffner et al., 2009; Desta and Ortiz, 2014).

The tenfold cross validation prediction accuracy of rrBLUP results ranged 0.51–0.65, BayesA results ranged 0.46–0.69, BayesB results ranged 0.58–0.64, BayesCPi results ranged 0.54–0.68 and GBLUP results ranged 0.67–0.76 (Figure 4 and Table 3). The consistency of the prediction accuracy was better with BayesB and GBLUP across tenfold analysis but GBLUP stood out in comparison with rest of the models, it could be mainly because of great genetic relationship between training and validation set, GBLUP uses genetic relationship coefficients instead of marker effects for calculation of GEBVs of the lines. The average prediction accuracy across tenfold analysis ranged 0.60 (rrBLUP, BayesA, BayesB and BayesCPi) – 0.72 (GBLUP). Based on average values of tenfold cross validation across models, GBLUP stands out with highest accuracy. When large number of markers data was used with good genetic relationship between training and test set, BayesB and GBLUP models appears to be robust in comparison with rrBLUP, BayesA, BayesCPi, but one of the challenges could be computational power that can be further improved by using advanced statistical models

(Heffner et al. 2011b). In summary, the results indicated that GBLUP and BayesB were more efficient models to predict sheath blight resistance.
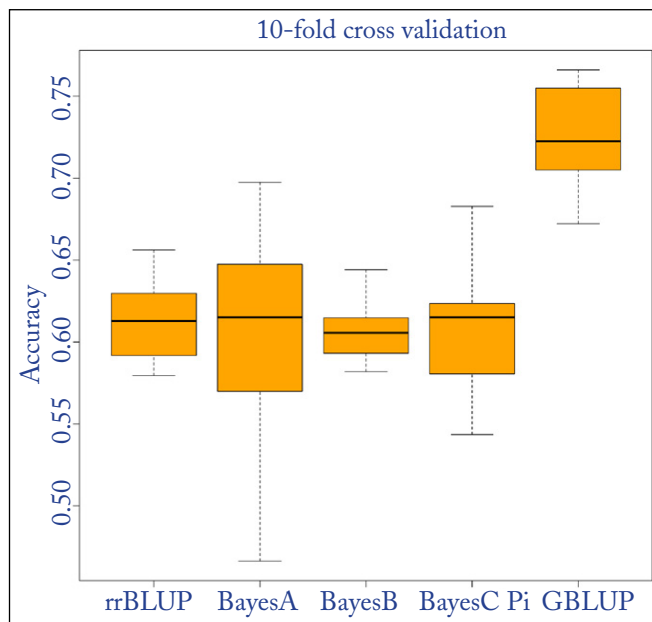

Figure 4: Box plot of tenfold cross validation analysis results of all five genomic selection models studied

3.1. 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 and GBLUP models. The analysis was done keeping the training and validation set size constant (1390 and 155 lines respectively). The average prediction accuracy with Bayes B model 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 5. The results in Table 4 clearly reveals that there was high range of prediction accuracy values across tenfold especially for lower MD datasets and accuracy was quite consistent with higher MD datasets.
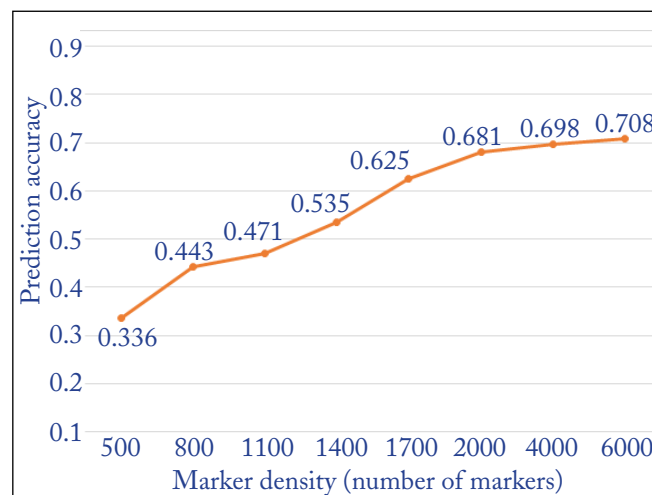

Figure 5: Line graph depicting average prediction accuracies of different levels of marker densities studied with Bayes B model

Whereas, the average prediction accuracy with GBLUP model across ten-fold cross validation obtained was 0.432 with MD=500, 0.539 with MD=800, 0.562 with MD=1100, 0.632 with MD=1400, 0.722 with MD=1700, 0.779 with MD=2000, 0.784 with MD=4000, and 0.793 with MD=6000. Prediction accuracy enhanced as the MD improved, a strong response to increase in marker density up to 2000 markers was witnessed with only a negligible increase in prediction accuracy when increased from 2000

Table 3: Tenfold cross validation analysis results of all five genomic selection models studied

Tenfold	rrBLUP	BayesA	BayesB	BayesCPi	GBLUP
1	0.596052	0.541081	0.581831	0.580503	0.704909
2	0.62977	0.667377	0.644163	0.545643	0.707266
3	0.511066	0.614166	0.593093	0.581049	0.723661
4	0.626937	0.569762	0.606939	0.623068	0.766029
5	0.656228	0.612357	0.598611	0.682747	0.703411
6	0.591816	0.626874	0.609494	0.543562	0.672195
7	0.57953	0.466312	0.604107	0.613744	0.754901
8	0.612378	0.697367	0.614845	0.616392	0.721327
9	0.613143	0.615836	0.632851	0.623488	0.761969
10	0.629728	0.647432	0.590061	0.624802	0.734774
Average	0.604665	0.605856	0.6076	0.6035	0.725044

Table 4: Detailed information on prediction accuracy across tenfold for different density of markers used for analysis with Bayes B model

Fold	Marker density (number of markers)							
	500	800	1100	1400	1700	2000	4000	6000
1	0.385	0.454	0.493	0.495	0.582	0.693	0.683	0.713
2	0.313	0.452	0.475	0.495	0.577	0.698	0.694	0.707
3	0.396	0.412	0.462	0.517	0.645	0.695	0.663	0.713
4	0.245	0.425	0.495	0.563	0.682	0.662	0.708	0.685
5	0.389	0.473	0.478	0.567	0.672	0.672	0.713	0.713
6	0.282	0.428	0.495	0.495	0.649	0.698	0.698	0.706
7	0.256	0.426	0.428	0.487	0.545	0.648	0.658	0.726
8	0.358	0.421	0.435	0.576	0.675	0.697	0.691	0.684
9	0.373	0.471	0.493	0.587	0.647	0.671	0.706	0.719
10	0.368	0.468	0.455	0.572	0.585	0.674	0.692	0.713
Average Accuracy	0.336	0.443	0.471	0.535	0.625	0.681	0.698	0.708

to 6000 markers, the trend of results was similar as that of Bayes B model but the prediction accuracy values were high revealing that GBLUP performance is better than Bayes B model. The results are summarized in Figure 6. The Table 5 indicates that there was high range among prediction accuracy values across tenfold especially for lower MD datasets and consistency improved with higher MD datasets.

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). The marker density at which plateau was reached in the current study was significantly higher than the plateau point of previous studies in smaller populations in wheat (Heffner et al., 2011b), as high marker
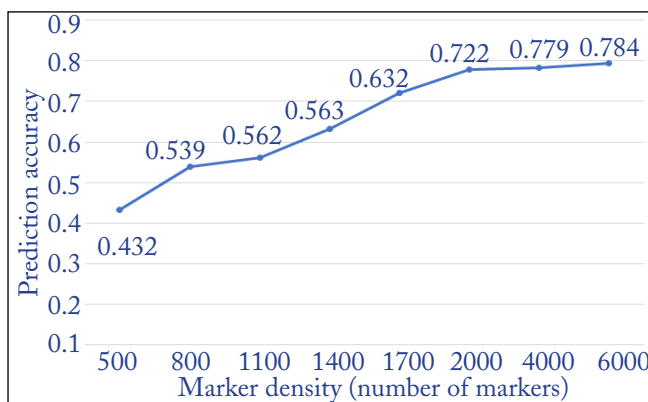

Figure 6: Line graph depicting average prediction accuracies of different levels of marker densities studied with GBLUP model

Table 5: Detailed information on prediction accuracy across tenfold for different density of markers used for analysis with GBLUP model

Fold	Marker density (number of markers)							
	500	800	1100	1400	1700	2000	4000	6000
1	0.495	0.554	0.583	0.605	0.702	0.803	0.783	0.803
2	0.402	0.532	0.565	0.575	0.667	0.788	0.774	0.797
3	0.506	0.532	0.552	0.637	0.755	0.815	0.773	0.803
4	0.325	0.505	0.575	0.643	0.762	0.742	0.788	0.765
5	0.519	0.603	0.608	0.697	0.802	0.802	0.843	0.813
6	0.342	0.488	0.545	0.565	0.699	0.768	0.758	0.806
7	0.354	0.524	0.526	0.585	0.643	0.746	0.756	0.804
8	0.434	0.498	0.513	0.651	0.749	0.775	0.762	0.762
9	0.478	0.579	0.598	0.69	0.752	0.772	0.811	0.774
10	0.472	0.575	0.557	0.674	0.69	0.779	0.794	0.811
Average Accuracy	0.43275	0.539	0.5622	0.6322	0.7221	0.779	0.7842	0.7938

densities only facilitate finer resolution and more accurate estimates of QTL effects when combined with large population size and low linkage disequilibrium (Huang et al., 2012). This analysis showed that response to increased marker density is largest when using a diverse training set to predict between poorly related materials.

4. CONCLUSION

From the data fit and tenfold cross validation results it is evident that GBLUP and Bayes B models provide high prediction accuracy compared to other statistical models investigated in the present study. The study of effect of marker density on accuracy indicated that 2000 markers were enough for generating a relatively accurate prediction calibration for sheath blight. As the inheritance of sheath blight resistance is complex and also cost involved in genotyping has drastically reduced due to path breaking technologies in biotech industry, genomic selection can be promising strategy to tackle while breeding for sheath blight resistance in rice.

5. ACKNOWLEDGEMENT

6. REFERENCES

Amandeep, K., Dhaliwal, L.K., Pannu, P.P.S., 2015. Role of meteorological parameters on sheath blight of rice under different planting methods. International Journal of Bio-resource and Stress Management 6(2), 214–219.

Bernardo, R., Yu, J., 2007. Prospects for genome wide selection for quantitative traits in maize. Crop Science 47, 1082–1090.

Chen, Z., Feng, Z., Kang, H., Zhao, J., 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.

Desta, Z., Ortiz, R., 2014 Genomic selection: genome-wide prediction in plant improvement. Trends Plant Science 19, 592–601.

Heffner, E., Sorrells, M., Jannink, J., 2009 Genomic selection for crop improvement. Crop Science 49, 1–12.

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

Heffner, E.L., Jannink, J., Iwata, J.H., Souza, E., Sorrells, M.E., 2011b. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Science 51, 2597–2606.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.L., Sorrells, M.E., 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9), e74612.

Huang, B., George, A., Forrest, K., Kilian, A., Hayden, M., 2012. A multiparent advanced generation inter-cross population for genetic analysis in wheat. Journal of Plant Biotechnology 10, 826–839.

Isidro, J., Jannink, J., Akdemir, D., Poland, J., Heslot, N., 2015. Training set optimization under population structure in genomic selection. Theoretical and Applied Genetics 128, 145–158.

Lorenz, A.J., Smith, K.P., Jannink, J.L., 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Science 52, 1609–1621.

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

Molla, K.A., Karmakar, S., Molla, J., Bajaj, P., Varshney, R.K., Datta, S.K., 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., 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 Rhizoctonia solani AG1-IA. Journal of Fungi 6, 71–82.

Rutkoski, J.E., Heffner, E.L., Sorrells, M.E., 2011. Genomic selection for durable stem rust resistance in wheat. Euphytica 179, 161–173.

Rutkoski, J., Benson, J., Jia, Y., Brown Guedira, G., Jannink, J.L., Sorrells, M., 2012. Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. Plant Genome 5(2), 51.

Rutkoski, J.E., Poland, J.A., Singh, R.P., Huerta Espino, J., Bhavani, S., Barbier, H., Matthew, N.R., Jannink, J.L., Sorrells, M.E., 2014. Genomic selection for quantitative adult plant stem rust resistance in wheat. Plant Genome 7(3), 34–46.

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

Yellareddygari, S.K.R., Reddy, M.S., Kloepper, J.W., Lawrence, K.S., Fadamiro, H., 2014. Rice sheath blight: a review of disease and pathogen management approaches. Journal of Plant Pathology and Microbiology 5, 4–22.

Zhang, S.W., Yang, Y., 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., Pan, X.B., 2010. Current progress on genetics and breeding in resistance to rice sheath blight. Scientia SinicaVitae 40, 1014–1023.

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