Original Research Article

Validation of genomic selection approach for predicting sheath blight resistance in Rice (*Oryza sativa* L.,)

ABSTRACT

Rice sheath blight (ShB) is one of the most serious fungal diseases caused by *Rhizoctonia solani*, instigating significant yield losses in many rice-growing regions of the world. Intensive studies indicated that resistance for sheath blight is controlled possibly by polygenes. Because of complex inheritance, it's very difficult to exploit and tap all the genomic regions conferring resistance using classical approaches of QTL mapping, it's very important to have a different strategy to harness such resistance mechanism. One promising approach that can potentially provide accurate predictions of the resistance phenotypes is genomic selection (GS). The research was undertaken with an objective to validate genomic selection approach for predicting sheath blight resistance involving 1545 Recombinant inbred lines (RILs) derived from eleven crosses between resistant and susceptible parents (Jasmine 85XTN1, Jasmine 85XSwarnaSub1, Jasmine 85XII32B, Jasmine 85XIR54, TetepXTN1, TetepXSwarna Sub1, TetepXII32B, TetepXIR54, MTU 9992XTN1, MTU 9992XII32B and MTU 9992XIRBB4). Where, Jasmine 85, Tetep & MTU 9992 were resistant parents and TN1, Swarna Sub1, II32B, IR54 & IRBB4 were susceptible parents. During rainy season (2020) the F_7 RILs were screened for their reaction to sheath blight in two hot spot locations. The genotyping was done with Illumina platform having 6564 SNP markers. Bayesian B approach was used to train the statistical model for calculation of marker effects and GEBVs. The data fit prediction accuracy of training set turned out to be 0.70 and random cross validation with different approaches the prediction accuracy ranged from 0.67 to 0.74. The results are lucrative, all in all, high prediction accuracies observed in this study suggest genomic selection as a very promising breeding strategy for predicting sheath blight resistance in Rice.

Keywords: *Rice, sheath blight, SNPs, genomic selection, BayesB*

INTRODUCTION

Rice (*Oryza sativa* L.) feeds more than half of the world's population and genetic improvement of this food crop can serve as a major component of sustainable food production.

Rice sheath blight (ShB) is one of the most devastating fungal diseases of rice, causing significant yield losses in many rice-growing regions of the world. This disease has become popular recently because of intensification of rice-cropping systems with the development of new short stature, high tillering, high yielding cultivars, high plant densities, and an increase in nitrogen fertilization, these morphological and microenvironment situations are very much congenial for the growth and multiplication of the sheath blight fungus. In India its prevalence is mainly confined to coastal places of India where farmers grow very high yielding varieties and hot humid climate adds to that. These factors promote disease spread by providing a favorable microclimate for the disease agent due to a dense leaf canopy with an increased leaf-to-leaf and leaf-to-sheath contact (Banniza *et al*., 2007).

The necrotrophic Sheath Blight pathogen possess a broad range of hosts, there are few germplasm lines which are known to show resistant reaction against this pathogen, most of the breeders are focused on harnessing these resistant sources to breed cultivars which are resistant to tolerant for this disease.

Because of lack of availability of more number of authentic and reliable sources of resistance, breeding for sheath blight has been challenging in Rice (Jia *et al*. 2009; Zuo *et al*. 2010; Srinivasachary, Willocquet and Savary 2011). Upon intensive study, it's believed to be controlled by many genomic regions dispersed across the genome (Zuo *et al*. 2013). It is widely thought that quantitative nature of resistance could be the expedient for evolving varieties with durable/horizontal resistance (Poland *et al*. 2013) against sheath blight in Rice.

Breeding for disease resistance involving many genes is a challenge. One of popular approach which is very popular now a days which can help in breeding for complex traits is genomic selection, because of reduction in cost involved in genotyping and development of robust statistical models this approach is becoming very popular while breeding for quantitative resistance now a days. Genomic selection uses large number of markers scattered across the genome to obtain the genomic estimated breeding values of individuals (Meuwissen *et al*., 2001). It was shown to be especially 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, 2014).

In general, prediction accuracies are reported to be higher in bi-parental populations than in populations of complex crosses with good genetic relationship between training and test population (Bernardo & Yu, 2007). The current investigation was done with eleven bi-parental populations created by design to validate the efficiency of genomic selection approach to predict sheath blight resistance with different random cross validation methods.

MATERIAL AND METHODS

Parent material and phenotyping of F_7 RILs for ShB

A total of 250 germplasm lines were screened for identification of lines which were resistant and susceptible for Sheath blight. Crosses were made involving Jasmine 85, Tetep & MTU 9992 as resistant to moderately resistant parents and TN1, Swarna Sub1, II32B, IR54 & IRBB4 as susceptible parents. The total of 1545 RILs across eleven 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 85XTN1, Jasmine 85XSwarnaSub1, Jasmine 85XII32B, Jasmine 85XIR54, TetepXTN1, TetepXSwarna Sub1, TetepXII32B, TetepXIR54, MTU 9992XTN1, MTU 9992XII32B and MTU 9992XIRBB4. 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 16008’ N and Longitude 81008’ E, Latitude 17°10’N and Longitude 81°41’ 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 with row-to-row distance 15 cm and plant to plant distance 10 cm 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 so 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 five phenotypic categories, where score 1: 0-20%, score 3: 21-30%, score 5: 31-45%, score 7: 46-65% and score 9: 66-100%.

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.

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 provides information about data fit 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 according to the marker effects estimated in the training set.

For the current investigation, Bayesian B model was used for training the model and to generate marker effects to get GEBV's of the breeding lines.

The Bayesian models assume a prior marker effects distribution and are of the form:

$$y = 1_n\mu + X\beta + \varepsilon$$

Where X is the incidence matrix for the markers and β is the vector of k marker effects. The Bayesian B model was implemented in GATK (Genome Analysis Tool Kit) tool with background of 'R' package with 50,000 iterations.

Bayes B, proposed by Meuwissen *et al.* (2001) uses a mixture distribution prior where marker effects are assumed to be zero with probability, π and marker effects are assumed to be drawn from a scaled-t distribution with probability, 1-π. In BA, π = 0, but BB assumes that many markers have no effect at all and hence π > 0 (Habier *et al*., 2011). Heffner et al. (2011) referred to this as a more realistic prior because certain regions of the genome are expected to have no quantitative trait loci (QTL) and thereby zero effect. The tool treats the parameter p (proportion of non-null effects) as unknown and assigns a Beta (B) prior parametrized such that the expected value by $E(\pi) = \pi_o$ and p_o is the number of prior counts. The prior densities for BB is represented as

$$p(\boldsymbol{\beta}_j, \sigma^2_{\beta}, \pi) = \{\Pi_k[\pi N(\boldsymbol{\beta}_{jk}|0, \sigma^2_{\beta}) + (1-\pi)1(\boldsymbol{\beta}_{jk} = 0)]\chi^{-2}(\sigma^2_{\beta jk}|df_{\beta}, S_{\beta})\}\ B(\pi|p_0, \pi_0) \times G(S_{\beta}|r, s)$$

Cross validation analysis

The cross-validation study was done with two approaches. In first approach, all the lines were randomly divided into training set and validation set with different percentage of individuals in each set, using phenotypic and genotypic data of all the individuals which were part of training set, the marker effects were created, the validation set individuals GEBVs were calculated by summing all marker effects taken from training set using only genotypic data, later predicted values (GEBVs) were correlated with phenotypic values to know prediction accuracy of training set, the analysis was run ten times. Finally, prediction accuracies across ten replications were averaged.

Whereas, in second approach (family drop method), out of eleven populations, ten populations individuals were made part of training set and individuals of one population were made part of validation set, ensured that every population will be part of validation set at least once. Using phenotypic and genotypic data of all the individuals which are part of training set, the marker effects were generated, the validation set individuals GEBVs were calculated by summing all marker effects taken from training set using only genotypic data, later predicted values (GEBVs) were correlated with phenotypic values to get the prediction accuracy, the analysis was run ten times, finally the prediction accuracy values across ten runs were averaged.

RESULTS AND DISCUSSION

The frequency distribution of 1545 F_7 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 9992XTN1 and MTU 9992XIRBB4 possessed lowest and highest number of informative markers (Table 2). Marker effects generated after statistical analysis (Figure 3) clearly explained that several loci scattered across the genome were contributing to sheath blight resistance, which demonstrated that the resistance to sheath blight was governed by many genes with additive effect, this was reported by earlier researcher (Zuo *et al*. 2013).

The prediction accuracy of training set (data fit) across all populations studied turned out to be 0.70 (Figure 4). The data fit analysis was done by population as well, among the different populations studied the prediction accuracy ranged from 0.18 to 0.67, MTU 9992XIRBB4 and TetepXSwarna Sub1 exhibited lowest and highest prediction accuracy respectively (Figure 5). When large number of markers data is available Bayes B model appears to be robust in comparison with ridge regression, Bayes A, Bayes C, stepwise regression etc., but one of the challenges could be computational power that can be improved by using advanced statistical models (Heffner *et a*l. 2011). The prediction accuracy of training set also depends on many factors, like phenotypic precision, size of the training set, number of markers used, LD between markers and traits of interest, statistical model used, marker type, heritability of the trait, genetic relationship between training and test set etc.

Results of first cross validation approach, in which different percentages of individuals were made part of training and validation set, the average prediction accuracy with three different population sizes ranged from 0.67 to 0.74, lowest was observed when only 25 percentage of the individuals were part of training set and highest when 75 percentage of individuals were part of training set (Table 3, 4 and 5), this clearly showed that the number of individuals in training set would have great impact on prediction accuracy. Results of second method of cross validation (population drop approach) illustrated prediction accuracy range of 0.64 to 0.72 (Table 6) across ten replications with a unique population being part of validation set in each replication, average prediction accuracy realized was 0.69 (Table 6). The data fit and random cross validation results of training set were appealing from both the approaches, which exemplified that training set developed with Bayes B model possess good predictability.

CONCLUSION

From the data fit and cross validation results it is evident that genomic selection method can be successfully used for predicting sheath blight resistance in Rice. The training set developed with eleven populations can be used further to predict untested populations. As cost involved in genotyping has drastically reduced due to path breaking technologies in biotech industry, genomic selection can be successfully and efficiently implemented to tackle the complex traits like sheath blight to increase the rate of genetic gain in Rice breeding.

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

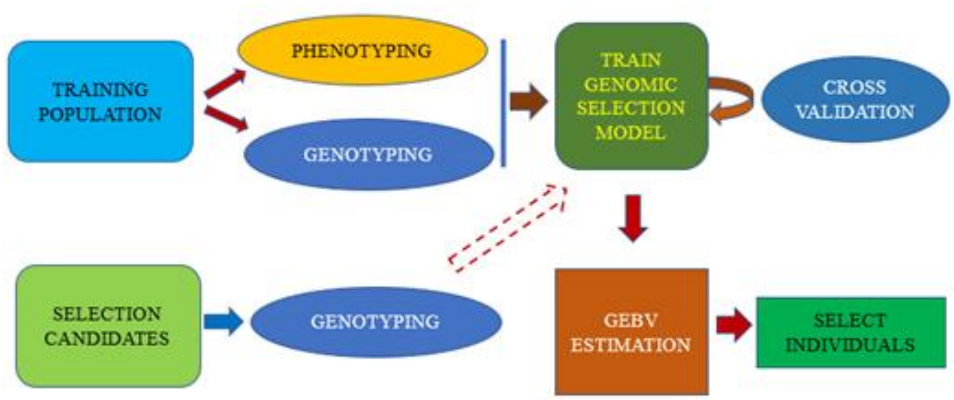

Figure 2: Distribution of sheath blight phenotypic scores into five classes or categories

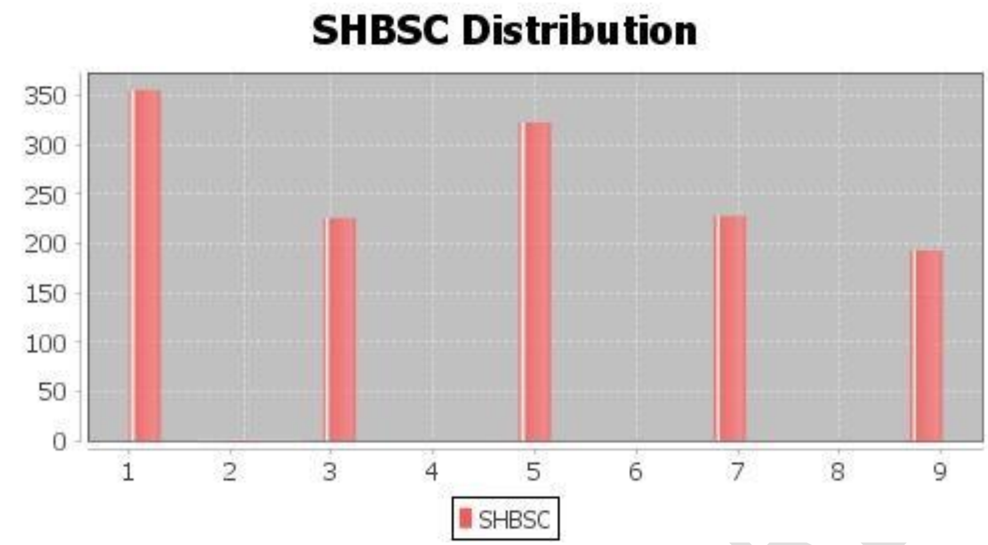

Figure 3: Marker effects of loci spread across the genome associated with sheath blight generated by marker trait association using Bayesian B model

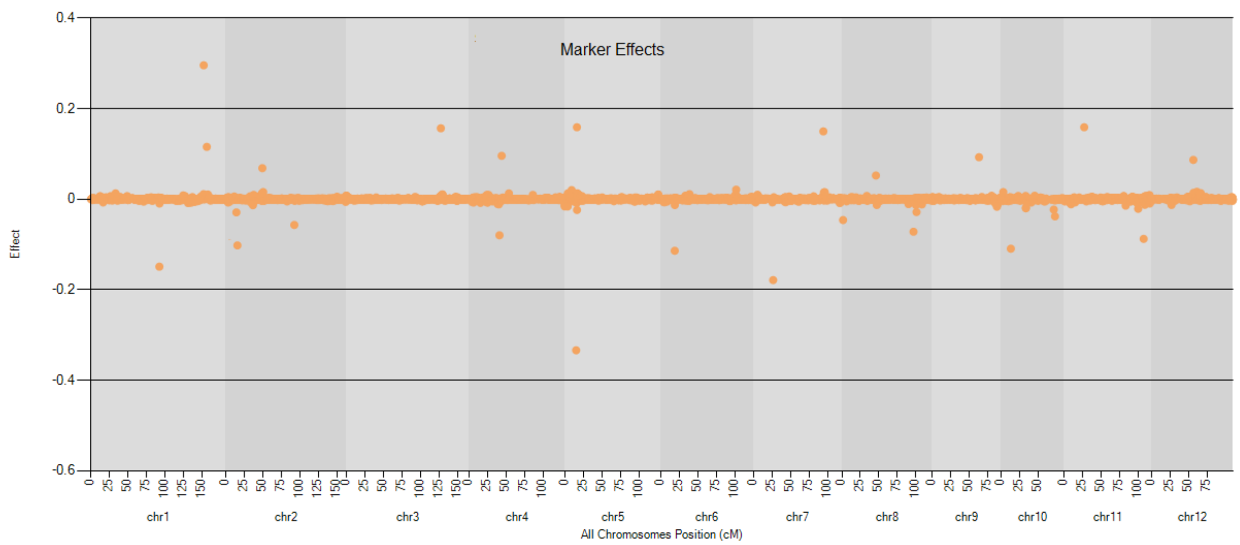

Figure 4: Scatter plot showing data fit analysis of training set across eleven populations studied (correlations between phenotypic values and genomic estimated breeding values of same data set for sheath blight scores).

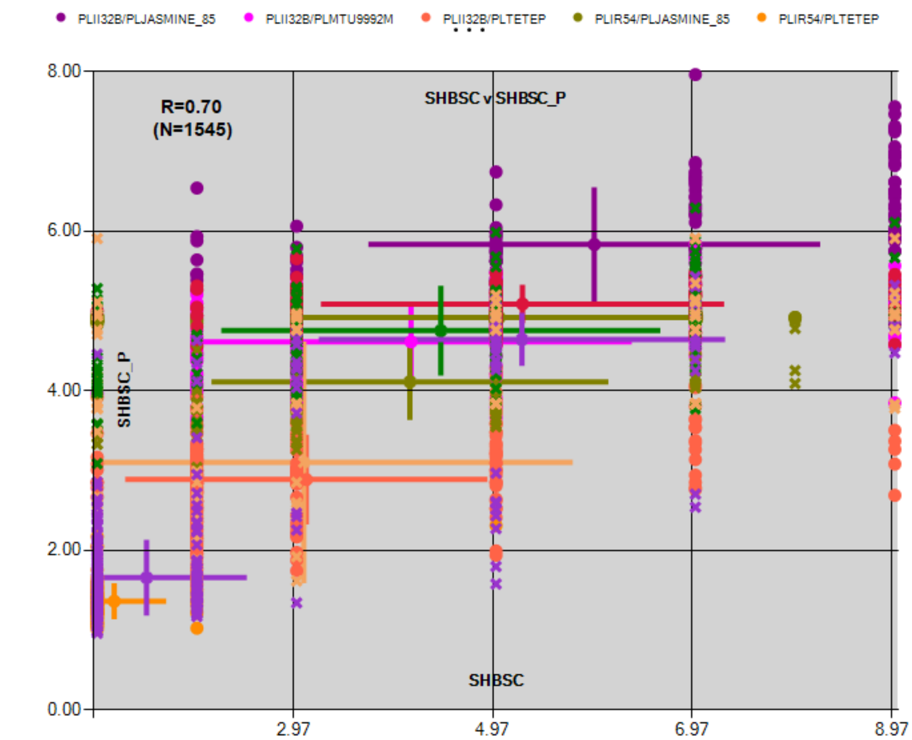

Figure 5: Histogram showing data fit analysis by population (correlations between phenotypic values and genomic estimated breeding values of same data set for sheath blight scores).

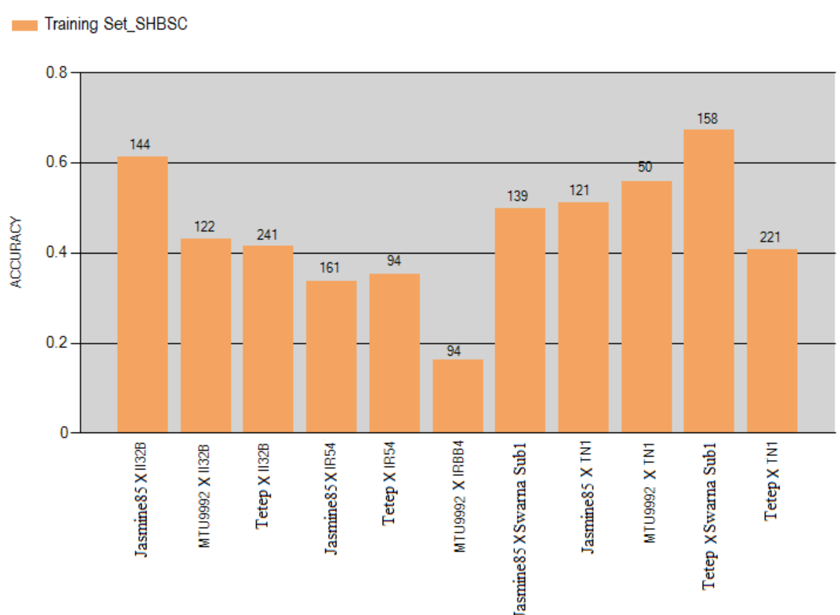

Table 1: Summary of marker data used for analysis and SNPs distribution on each chromosome

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	Total Markers	Polymorphic Markers
Jasmine 85XTN1	6564	2522
Jasmine 85XSwarna Sub1	6564	2627
Jasmine 85XII32B	6564	2586
Jasmine 85XIR54	6564	2663
TetepXTN1	6564	2806
TetepXSwarna Sub1	6564	2278
TetepXII32B	6564	2702
TetepXIR54	6564	2796
MTU 9992XTN1	6564	1407
MTU 9992XII32B	6564	2314
MTU 9992XIRBB4	6564	2849

Table 3: Cross validation analysis results with the training set possessing 75% of the individuals and validation set with 25% of the individuals

Random Cross Validation of Training Set : Validation Set (75:25 Percentage)

Trait	Replication	Total #obs	Training Set #obs	Validation Set #obs	Accuracy
SHBSC	Rep01	1545	1157	388	0.6945
SHBSC	Rep02	1545	1157	388	0.7403
SHBSC	Rep03	1545	1157	388	0.6925
SHBSC	Rep04	1545	1157	388	0.7621
SHBSC	Rep05	1545	1157	388	0.7475
SHBSC	Rep06	1545	1157	388	0.801
SHBSC	Rep07	1545	1157	388	0.7294
SHBSC	Rep08	1545	1157	388	0.7538
SHBSC	Rep09	1545	1157	388	0.7604
SHBSC	Rep10	1545	1157	388	0.7483
				Average	**0.74298**

Table 4: Cross validation analysis results with the training set possessing 50% of the individuals and validation set with 50% of the individuals

Random Cross Validation of Training Set : Validation Set (50:50 Percentage)					
Trait	**Replication**	**Total #obs**	**Training Set #Obs**	**Validation Set #Obs**	**Accuracy**
SHBSC	Rep01	1545	773	772	0.7269
SHBSC	Rep02	1545	773	772	0.7143
SHBSC	Rep03	1545	773	772	0.725
SHBSC	Rep04	1545	773	772	0.6978
SHBSC	Rep05	1545	773	772	0.7503
SHBSC	Rep06	1545	773	772	0.7456
SHBSC	Rep07	1545	773	772	0.6952
SHBSC	Rep08	1545	773	772	0.7059
SHBSC	Rep09	1545	773	772	0.7005
SHBSC	Rep10	1545	773	772	0.7081
				Average	**0.71696**

Table 5: Cross validation analysis results with the training set possessing 25% of the individuals and validation set with 75% of the individuals

Random Cross Validation of Training Set : Validation Set (25:75 Percentage)					
Trait	**Replication**	**Total #obs**	**Training Set #obs**	**Validation Set #obs**	**Accuracy**
SHBSC	Rep01	1545	386	1159	0.6714
SHBSC	Rep02	1545	386	1159	0.7042
SHBSC	Rep03	1545	386	1159	0.6875
SHBSC	Rep04	1545	386	1159	0.6677
SHBSC	Rep05	1545	386	1159	0.6602
SHBSC	Rep06	1545	386	1159	0.6808
SHBSC	Rep07	1545	386	1159	0.6526
SHBSC	Rep08	1545	386	1159	0.6738
SHBSC	Rep09	1545	386	1159	0.6993
SHBSC	Rep10	1545	386	1159	0.6824
				Average	**0.67799**

Table 6: Cross validation analysis results with the training set possessing the individuals of ten families and validation set with the individuals of one family

Random Cross Validation of Training Set_Family drop method						
Trait	**Total #Obs**	**Training Set_Families**	**ValidationSet_Family**	**Training Set #Obs**	**Validation Set #Obs**	**Accuracy**
SHBSC	1545	Ten Families	Jasmine 85XTN1	1424	121	0.7183
SHBSC	1545	Ten Families	Jasmine 85XSwarna Sub1	1406	139	0.7176
SHBSC	1545	Ten Families	Jasmine 85XII32B	1401	144	0.6647
SHBSC	1545	Ten Families	Jasmine 85XIR54	1384	161	0.7164
SHBSC	1545	Ten Families	TetepXTN1	1324	221	0.6459
SHBSC	1545	Ten Families	TetepXSwarna Sub1	1387	158	0.7054
SHBSC	1545	Ten Families	TetepXII32B	1304	241	0.7207
SHBSC	1545	Ten Families	TetepXIR54	1451	94	0.6679
SHBSC	1545	Ten Families	MTU 9992XTN1	1495	50	0.7057
SHBSC	1545	Ten Families	MTU 9992XII32B	1423	122	0.7181
SHBSC	1545	Ten Families	MTU 9992XIRBB4	1451	94	0.7036
					Average	**0.69857**

COMPETING INTERESTS DISCLAIMER:

Authors have declared that no competing interests exist. The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

REFERENCES

Banniza S, A. A. Sy, P. D. Bridge, S. A. Simons, and M. Holderness (2007) Characterization of Populations of *Rhizoctonia solani* in Paddy Rice Fields in Côte d'Ivoire. Published Online:22 Feb 2007https://doi.org/10.1094/PHYTO.1999.89.5.414

Bernardo R and Yu J (2007) Prospects for Genome wide Selection for Quantitative Traits in Maize. Crop Science 47: 1082-1090.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

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

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

Heslot, N., J.L. Jannink, and M.E. Sorrells. 2013b. Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci 53(June): 921–933.

Heslot N. et al (2012) Genomic selection in plant breeding: a comparison of models. *Crop Science* 52: 146-160.

Heslot, N., H.-P. Yang, M.E. Sorrells, and J.-L. Jannink. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52: 146–160.

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

Jia Y, Liu GJ, Costanzo S, Lee SH, Dai YT (2009) Current progress on genetic interactions of rice with rice blast and sheath blight fungi. Front Agri in China 3:231–239.

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

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

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink, and M. Sorrells. 2012. Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. Plant Genome J 5(2): 51.

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

Rutkoski, J.E., J.A. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, H. Barbier, et al. 2014. Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. Plant 96 Genome 7(3).

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.

Zuo SM, Zhang YF, Chen ZX, Chen XJ, Pan XB (2010) Current progress on genetics and breeding in resistance to rice sheath blight. Scientia Sin Vitae 40:1014–1023.

Zuo SM, Yin YJ, Zhang L, Zhang YF, Chen ZX, Pan XB (2013) Fine mapping of qSB-11LE, the QTL that confers partial resistance on rice sheath blight. Theor Appl Genet 126:1257–1272.