bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Genome Wide Association Studies on 7 Yield-related Traits of 183 Rice Varieties in Bangladesh

Nilanjan Roy[1], Acramul Haque Kabir[1], Nourin Zahan[1], Shahba Tasmiya Mouna[1], Sakshar Chakravarty[2], Atif Hasan Rahman[2] and Md. Shamsuzzoha Bayzid[2,*]

[1]Department of Biomedical Engineering
Military Institute of Science and Technology
Dhaka-1216, Bangladesh
[2]Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
[*]Corresponding author: shams_bayzid@cse.buet.ac.bd

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Genome Wide Association Studies on 7 Yield-related Traits of 183 Rice Varieties in Bangladesh

Abstract

Rice genetic diversity is regulated by multiple genes and is largely dependent on various environmental factors. Uncovering the genetic variations associated with the diversity in rice populations is the key to breed stable and high yielding rice varieties. We performed Genome Wide Association Studies (GWAS) on 7 rice yielding traits (grain length, grain width, grain weight, panicle length, leaf length, leaf width and leaf angle) based on 39,40,165 single nucleotide polymorphisms (SNPs) in a population of 183 rice landraces of Bangladesh. Our studies reveal various chromosomal regions that are significantly associated with different traits in Bangladeshi rice varieties. We also identified various candidate genes, which are associated with these traits. This study reveals multiple candidate genes within short intervals. We also identified SNP loci, which are significantly associated with multiple yield-related traits. The results of these association studies support previous findings as well as provide additional insights into the genetic diversity of rice. This is the first known GWAS study on various yield-related traits in the varieties of *Oryza sativa* available in Bangladesh – the fourth largest rice producing country. We believe this study will accelerate rice genetics research and breeding stable high-yielding rice in Bangladesh.

Keywords: Genome Wide Association Studies (GWAS); Rice; yield-related traits; Single Nucleotide Polymorphism (SNP).

1 Introduction

Rice (*Oryza sativa* L.) is one of the most important food crops and feeds half the world's population. Especially, this is the staple food of about 160 million people in Bangladesh, and this country has one of the highest per capita consumption of rice. Therefore, its food security largely depends on the good harvest of rice. Future increases in rice production, required to feed a continuously growing population of this country amidst various adverse climatic conditions due to the climate change and limited arable land resources will rely primarily on genetic improvement of rice cultivars. Therefore, understanding the genetic basis of physiological and morphological variation in rice landraces in Bangladesh is critical for improving the quality and quantity of rice production. During the last few decades, great efforts in rice research have been made by Bangladesh Rice Research Institute (BRRI), in association with International Rice Research Institute (IRRI) to boost rice production. However, the current effort in increasing rice production in Bangladesh

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

is mostly based on analyzing morphological characteristics and developing hybrids with *trial-and-error*. This traditional approach is not "scalable" to investigate the tremendous genetic and phenotypic variation of thousands of rice varieties available in Bangladesh. GWAS, considering the rice varieties in Bangladesh, may reveal important genotype-phenotype associations which will direct the agricultural scientists towards a more informed research for breeding better rice varieties with desirable phenotypes suitable for the climate of Bangladesh.

Genome wide association studies (GWAS) have become a popular method to classify advantageous alleles and quantitative trait loci (QTL) associated with large-scale complex traits in rice population. Due to the growing awareness of the efficacy of GWAS in molecular dissection of traits and the abundance of genomic and phenotypic resources, many GWAS studies have been conducted over the past few years on various rice varieties across the world. Huang *et al.* (2010) [1] performed an association study on 14 rice agronomic traits across 373 indica rice varieties, and identified a total of 80 related sites. Huang *et al.* (2012) [2] identified a total of 32 heading date sites and 20 grain type sites across 950 rice varieties. Zhao *et al.* (2011) performed a GWAS on 34 traits across 413 rice varieties from 82 countries, and identified 234 associated sites [3]. Ya-fang *et al.* (2014) analyzed 315 rice varieties from the International Core Rice Germplasm Bank to perform a GWAS on five panicle traits, and a total of 36 candidate associated regions were detected [4]. Yang *et al.* (2014) performed GWAS on 15 traits, including 13 traditional agronomic traits and identified 141 associated loci [5]. Then they compared how these traits change along with the ecological environment. This led to the identification of valuable varieties and sub-groups with more favorable alleles. Biscarini *et al.* (2016) [6] conducted a genome-wide association analysis for grain morphology and root architecture for temperate rice accessions adapted to European pedo-climatic conditions, and a set of 391 rice accessions were GBS-genotyped leading to 57,000 polymorphic and informative SNPs. Among which 54% were in genic regions. A total of 42 significant genotype-phenotype associations were detected: 21 for plant morphology traits, 11 for grain quality traits, 10 for root architecture traits. The results helped them to dig into the narrow genetic pool of European temperate rice and to identify the most relevant genetic components contributing to a high yield of this germplasm. Zhang *et al.* (2019) [7] performed a GWAS with EMMAX for 12 agronomic traits using Ting's core collection (7128 rice landraces from all over China and from some of the other main rice-cultivating countries collected by Ying Ting [8]). Yang *et al.* (2019) [9] detected SNP loci and determined related genes affecting the rice grain shape which lead to high-yielding breeding of rice. In that study, a total of 161 natural Indica rice varieties grown in southern China were used for a GWAS of grain shape-related traits. These traits include grain length (GL), grain width (GW), 1000-grain weight (TGW), and grain length/width (GLW). Ma *et al.* (2019) conducted a GWAS and a gene called OsSNB was identified controlling the grain size in rice [10]. Similarly, significant efforts have been made for association mapping with other yield related traits (e.g., panicle and leaf traits) [4, 5, 7, 9, 11].

In this study, we performed genome-wide association studies on 7 yield traits across 183 rice varieties in Bangladesh. We leverage the 3K Rice Genome Project (3K RGP) [12], where 3,000 rice genomes were re-sequenced and a resulting set of over 19 million SNPs has been characterized and made accessible [12–14]. While the previous GWAS studies provide fundamental resources regarding association mapping on various rice traits, none

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of them were especially targeted for Bangladeshi rice varieties. However, the grain yield-related rice traits are regulated by multiple genes, which are significantly influenced by the environment [1,4,15,16]. As such, this study will further elucidate the impact of specific environmental conditions on the association between traits and genetic variations.

2 Materials and Methods

2.1 Rice Materials

We leveraged the data from 3K RGP [12] (snp-seek.irri.org), which has sequenced a core selection of 3,000 rice accessions from 89 countries. We filtered a total of 183 rice varieties of Bangladesh. Detailed information on these 183 rice varieties are provided in Supplementary Material SM2. We considered seven yield-related phenotypes, namely grain length (GL), grain width (GW), grain weight (GWT), panicle length (PL), leaf length (LL), leaf width (LW) and leaf angle (LA). Rice grain shapes are closely related to the yield and quality [9,17]. Leaf traits are among the major determinants of plant architecture, and are strongly associated to yield [18–21]. Panicle, being the top organ, is an important component in the canopy and is strongly correlated with spikelet yield [22]. Thus, investigating the genetic variations associated with these traits under specific conditions of Bangladesh would be fundamental to high-yield rice research in this country.

2.2 Data Analysis

The association analysis was done by Plink [23]. SNP markers with a missing rate of over 20 percent and minor allele frequency (MAF) less than 0.05 were removed to reduce false-positive rate. Thereby, we selected a total of 39,40,165 SNPs for conducting GWAS. We pruned our dataset of variants that are in linkage, and then performed Principle Component Analyses (PCA) to investigate as well as to correct for population structure by using the main components as covariates in the association tests. The quantitative association test function of Plink was used to get the P-values of significant SNPs. Significant threshold was set to $-\log_{10} P$, where $P = \frac{1}{n}$, $n =$ total number of markers used [7,9]. For the leaf angle, no association signal was higher than 6.58 (-$\log_{10} P$), and thus, we calculated another significant threshold based on the minimum Bayes factor (mBF), where mBF $= -eP \ln P$ [24]. Therefore, we used -$\log_{10}$mBF $= 4.98$ as the significant threshold for LA. A linkage disequilibrium threshold r^2 of 0.6 was used for the clumping of the independent significant SNPs. Physical distance threshold for clumping was set to the default value (250 Kb). Visualization and interpretation of the inferred association were done with the help of SNPEVG [25]. Candidate genes were screened through the Michigan State University (MSU) Rice Genome Annotation Project Database [26].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

3 Results and Discussion

3.1 Phenotypic diversity of the traits

Statistical analyses of the phenotypic diversity in the yield-related traits are shown in Table 1. The minimum and maximum GL in this population are 4.7 and 10.2, respectively, having a mean GL of 8.28 with 9.48% coefficient of variation. The CV values of various grain shapre related traits range from 9.48% ∼16.4%, and those of leaf and panicle traits range from 11.67% ∼57.7%. These analyses suggest that this group of rice varieties represent significant variations in grain shape and other yield-related traits.

The distribution of these traits as well the correlations between them are depicted in Fig. 1. It suggests that these traits are normally distributed, and GL, GW and GWT are positively correlated with each other, especially GW and GWT are strongly correlated with each other.

Table 1: **Statistical analyses of the yield-related traits.** We show the minimum (Min), maximum (Max), mean, standard deviation (SD) and coefficient of variation (CV) values for each of the 7 traits.

Trait	Min	Max	Mean	SD	CV (%)
GL (mm)	4.7	10.2	8.28	0.78	9.48
GW (mm)	2.1	3.9	3.05	0.33	10.8
GWT (gm)	1.2	3.5	2.4	0.4	16.4
LL (cm)	2	5	3.26	0.63	19.41
LW (cm)	0.8	2.3	1.29	0.24	19.26
LA (rd)	1	9	3.39	1.96	57.7
PL (cm)	18	32	24.89	2.90	11.67

3.2 Population structure

Principal component analysis was performed based on 39,40,165 SNPs in 183 rice varieties. PCA plot of the first two principal components is shown in Fig. 2 (a) which suggests three major subgroups: indica (IND), aus (AUS), aromatic (ARO). The principal components representing the IND samples are located in the lower left, the ARO samples are located in the upper right, and the AUS sample are located in the lower right part. There are a few ad-mixed (intermediate type) varieties as well. We also show the genome stratification based on SNP markers using multidimensional scaling (MDS) plot (see Fig. 2(b)).

Since population stratification is observed in the samples, we adjust for population stratification while testing for associations between the SNPs and the phenotypes. We find that the first two principal components capture the three main subgroups of rice varieties in our samples. So, we use the first two axes of variation as covariates in the P-value calculation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

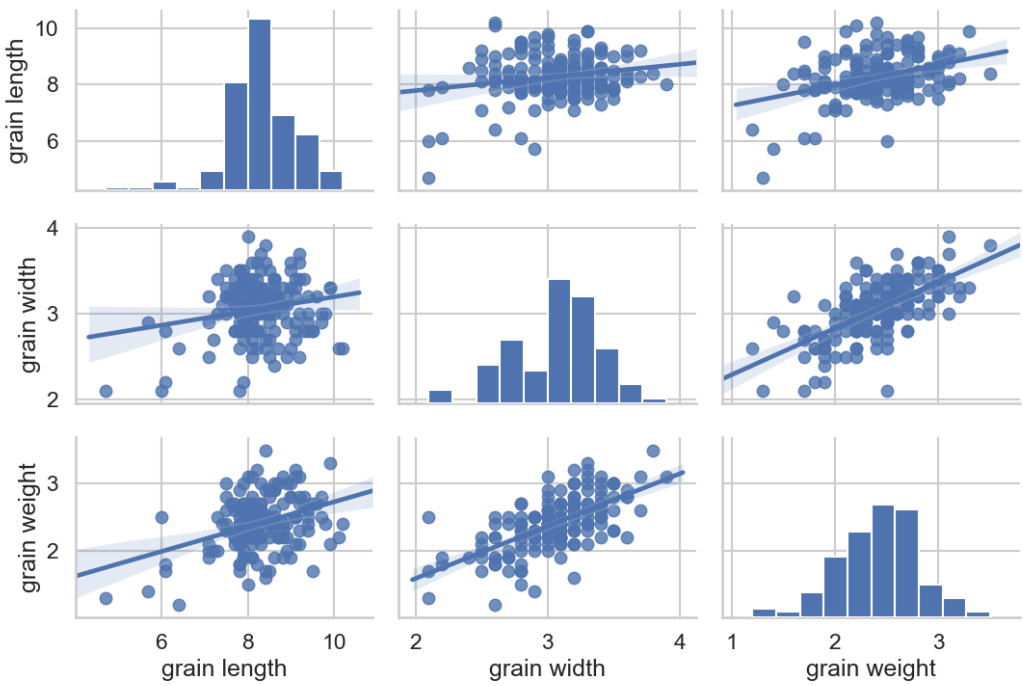
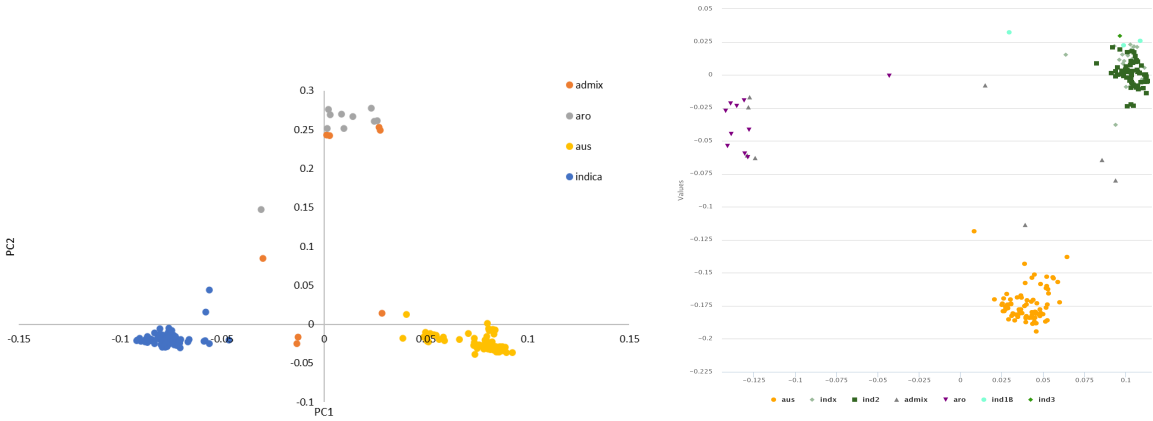


Figure 1: **Statistical analyses of the grain shape related traits.** We show the distribution and correlation scatter matrix of these three traits. GL and GW are shown in millimetre (mm) and GWT is shown in gram (gm). The blue lines in the scatter plots indicate the correlation trends.

3.3 Results on grain shape related traits

A total of 168 rice varieties (out of 183 Bangladeshi varieties collected from the 3K RGP) had GL, GW and GWT values associated with them, and we used these 168 varieties for our association studies. Manhattan plots (-$\log_{10}(P)$ genome-wide association plots) and quantile-quantile (QQ) plots for three grain shape related traits are shown in Fig. 3. We identified 15 lead SNP markers harbored in chromosomes 1, 3, 4, 5, 7, 8, 9, 10, and 12 that were significantly associated with GL (Fig. 3(a)). Both chromosomes 4 and 5 harbor three peak SNPs. The strongest peak was found in chromosome 5. The QQ plot (Fig. 3(d)) supports that some observed P-values are more significant than expected under the null hypothesis. The information regarding the peak SNPs are shown in Table 2. Nine candidate genes in chromosomes 1, 3, 4, 5, 7, and 10 were identified, harboring the SNPs which were associated with GL. Two candidate regions, chr04:4236793 and chr04:7788382, were found to be associated with a single candidate gene LOC_Os04g59624 (see Table 2). Notably, all three lead SNP loci in chromosome 5 (chr05_9776079, chr05_1025757, and chr05_23772096) were also found to be significantly associated with GWT (see Table 2). Moreover, SNP loci at chr05_23772096 – associated with the candidate gene LOC_Os05g40480 – was the strongest association peak for both GL and GWT. Thus, this study reveals chromosomal regions that are likely to regulate both GL and GWT, and thus stems the necessity of further future characterization. Furthermore, this loci was also significantly associated with panicle length (see Sec. 3.4 and Supplementary Materials SM1). However, future studies will need to identify and

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



(a) PCA plot (b) MDS plot

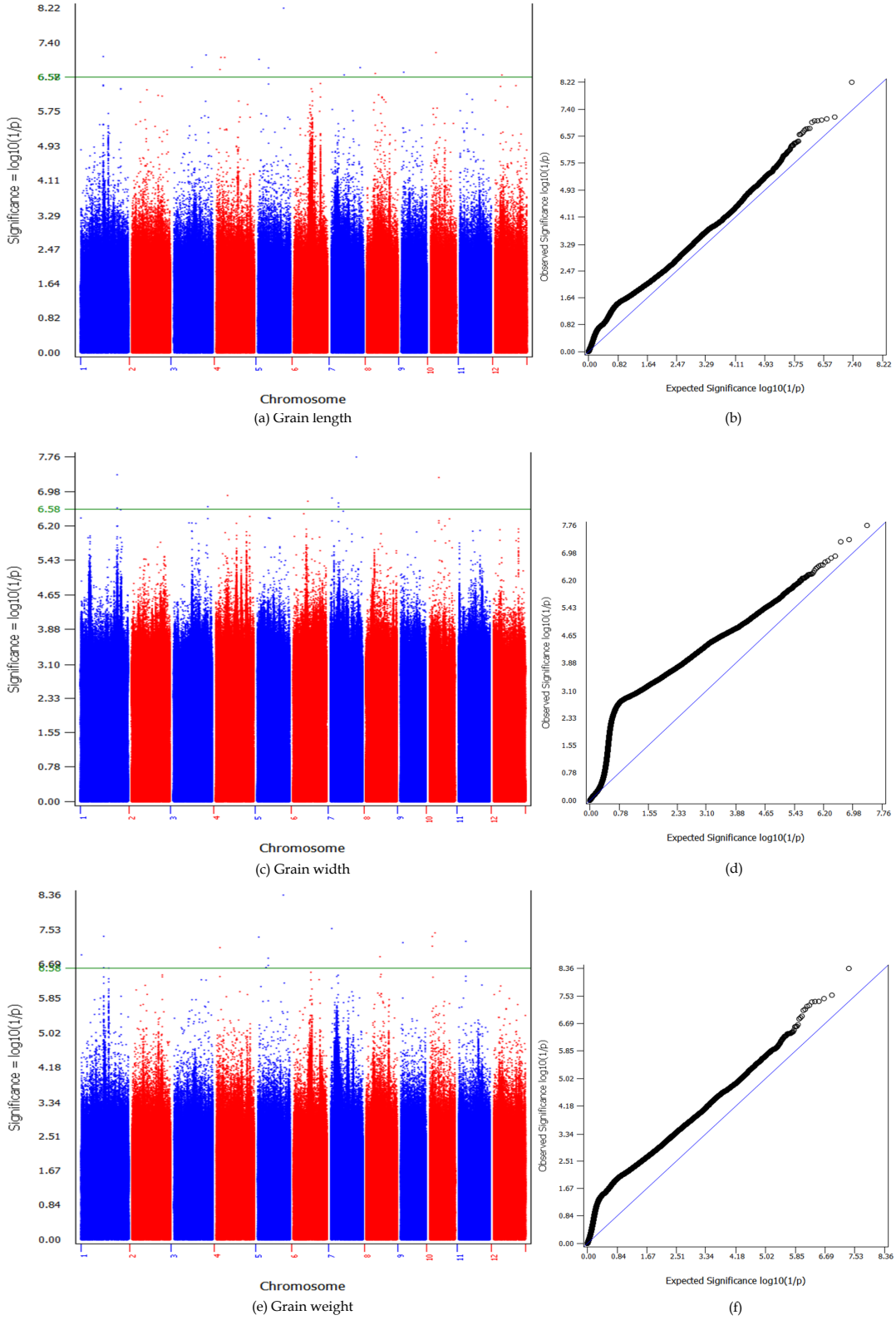
Figure 2: **Population structure using PCA and MDS.** (a) PCA plot of the first two principal components of 183 rice varieties. (b) MDS plot of the 183 rice varieties.

Our GWAS study supports previous findings as well as reveal additional chromosomal regions associated with GL. Yang *et al.* (2019) [9] identified eight SNP loci in four chromosomes (3, 5, 6, and 7) to be in close association with GL, and – among these four chromosomes – our study identified chromosomes 3, 5 and 7 to harbor seven significant loci. Ya-fang *et al.* (2015) identified 10 SNP loci in choromosomes 3, 5, 6, 7, 8, 10 and 12, which were significantly associated with GL [4]. A study on basmati rice of Indian origin (Singh *et al.* (2012) [27]) also revealed chromosomal regions in 1 and 7 to be associated with GL, and an SSR (simple sequence repeat) analysis in Pakistan (Aslam *et al.* (2014) [28]) identified chromosomal regions in 3 and 7 to be in close association with grain length. Kinoshita *et al.* [29] found chromosome 4 and 11 to harbor QTLs associated with GL.

Figure 3(c) shows the results on GW trait. Chromosomes 1, 3, 4, 6, 7, and 10 were found to harbor 9 lead SNPs, that were significantly associated with GW. These loci associated with GW included three strong peaks in chromosomes 7, 1 and 10, respectively. Notably, chromosome 7 included 4 strong peaks, including the strongest association signal associated with GW. Nine candidate genes were screened out based on these lead SNPs. Two lead SNP loci in chromosome 1 (chr01_33401190 and chr01_33410685) were associated with two candidate genes LOC_Os01g57760 and LOC_Os01g57770, and were within a short (10 kb) candidate interval. The presence of multiple candidate genes within such a short candidate interval suggest further investigation to explore the causality of genes may be beneficial, and subsequently may facilitate the identification of advantageous genetic variations. Ya-fang *et al.* (2015) [4] found chromosomes 1, 2, 3, 5, 6, 7, 10, 11 and 12 to harbor significant SNPs associated with GW. Yang *et al.* (2019) [9] identified various regions in chromosomes 4, 5, 6, 8, 9, and 11 to be associated with GW. Zhang *et al.* (2019) [7] discovered GS2 loci in chromosome 2 in close association with GW. Therefore, our studies have notable similarities as well as dissimilarities with prior studies.

Figure 3(e) shows the results on the GWAS for GWT. Chromosomal regions in 1, 4, 5, 7, 8, 9, 10 and 11 were identified to be associated with GWT that harbor 14

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



(a) Grain length (b)

(c) Grain width (d)

(e) Grain weight (f)

Figure 3: **GWAS on grain shape related traits using 168 rice varieties.** We show the Manhattan plots for GL, GW and GWT in figures (a), (c) and (e), respectively. The green line indicates the genome-wide significant threshold. Corresponding QQ-plots are shown in figures (b), (d) and (f). For QQ plots, the horizontal axis shows the $-\log10$-transformed expected P-values, and the vertical axis indicates $-\log10$-transformed observed P-values.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 2: **Genome-wide lead SNPs significantly associated with grain shape related traits.** For each trait, we show the significantly associated SNPs. Beta denotes the coefficient from a fit.

Trait	Chr	SNP ID	SNP site (bp)	Minor Allele	Major Allele	MAF	P	Beta (β)	Candidate gene
GL	5	174896782	23772096	A	G	0.09	6.05E-09	0.96	LOC_Os05g40480
	10	298342138	4856868	G	T	0.07	6.97E-08	1.03	LOC_Os10g01750
	3	109343151	30134978	T	C	0.06	7.93E-08	-0.96	LOC_Os03g52520
	1	20428432	20428432	T	A	0.08	8.64E-08	-0.76	LOC_Os01g36772
	4	119858785	4236793	T	C	0.05	9.08E-08	-1.09	LOC_Os04g59624
	4	123410374	7788382	T	G	0.07	9.11E-08	-0.92	LOC_Os04g59624
	5	152150443	1025757	A	G	0.12	1.01E-07	-0.83	LOC_Os05g02840
	3	96033585	16825412	A	G	0.06	1.55E-07	-1.05	LOC_Os03g29520
	7	238925722	26593815	G	A	0.06	1.59E-07	-1.08	LOC_Os07g44540
	5	160900765	9776079	A	G	0.12	1.62E-07	1.10	
	4	118951371	3329379	T	C	0.10	1.76E-07	0.90	
	9	272505994	2033444	G	C	0.05	2.04E-07	-4.87	
	8	249602184	7572656	C	T	0.10	2.19E-07	-0.97	
	7	224254658	11922751	G	A	0.10	2.38E-07	0.97	
	12	351735526	6021863	C	A	0.10	2.38E-07	0.91	
GW	7	236327454	23995547	A	G	0.10	1.75E-08	-0.26	LOC_Os07g40000
	1	33401190	33401190	C	T	0.22	4.40E-08	0.23	LOC_Os01g57760
	10	302089342	8604072	A	G	0.31	5.07E-08	-0.39	LOC_Os10g17140
	4	126597821	10975829	A	C	0.13	1.29E-07	2.72	LOC_Os04g19670
	7	213782817	1450910	G	A	0.07	1.47E-07	-0.29	LOC_Os04g03620
	6	194695024	13611904	A	G	0.14	1.74E-07	2.39	LOC_Os06g23300
	7	219851828	7519921	C	T	0.19	1.90E-07	1.96	LOC_Os07g13130
	3	111434100	32225927	T	C	0.49	2.30E-07	2.68	LOC_Os03g56572
	1	33410685	33410685	T	C	0.19	2.46E-07	1.92	LOC_Os01g57770
GWT	5	174896782	23772096	A	G	0.09	4.35E-09	0.50	LOC_Os05g40480
	7	213298898	966991	C	A	0.25	2.82E-08	-3.20	LOC_Os07g02674
	10	298342138	4856868	G	T	0.07	3.59E-08	0.52	LOC_Os10g08970
	1	20428432	20428432	T	A	0.08	4.36E-08	-0.39	LOC_Os01g36766
	10	295897683	2412413	G	T	0.08	4.40E-08	-2.67	LOC_Os10g04940
	5	152150443	1025757	A	G	0.12	4.54E-08	-0.44	LOC_Os05g02840
	11	323223710	6531153	C	T	0.37	5.76E-08	0.25	LOC_Os11g11760
	9	272505994	2033444	G	C	0.05	6.19E-08	-2.50	LOC_Os09g03960
	4	118951371	3329379	T	C	0.10	8.19E-08	0.46	
	1	161054	161054	C	A	0.09	1.23E-07	-0.46	
	8	254597466	12567938	A	G	0.06	1.36E-07	-2.29	
	5	160900765	9776079	A	G	0.12	1.48E-07	-2.93	
	5	158859162	7734476	T	C	0.07	2.50E-07	-2.28	
	1	25065564	25065564	A	G	0.37	2.58E-07	-0.21	

significant SNPs. Especially, chromosomes 1, 5, and 10 harbor more lead SNPs than others. We identified eight candidate genes in chromosomes 1, 5, 7, 9, 10 and 11, harboring the peak association signals associated with GWT. As we mentioned earlier, three lead SNP loci in chromosome 5 (chr05_9776079, chr05_1025757, and chr05_23772096) were significantly associated with both GWT and GL. As was observed for GL and GW, our association study on GWT presents some congruent results with respect to the previous studies [29, 30], in addition to revealing new chromosomal regions associated with GWT.

3.4 Results on Leaf and Panicle Traits

We performed association studies on four other yield related traits: panicle length and three leaf traits (leaf length (LL), leaf width (LW) and leaf angle (LA)). Among the 183 rice varieties in the 3K RGP project, four different subsets containing 158, 153, 160, 84 varieties had PL, LL, LW and LA trait values, respectively. The Manhattan plots on

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

these traits are shown in Fig. 4. Key observations include the identification of (i) a lead SNP in chromosome 5 which is also associated with GL, (ii) five lead SNPs (among 10) associated with leaf angle were found in chromosome 12, and four of them (chr12_3263244, chr12_3587574, chr12_3660122, and chr12_3650583) are within a short (400 kb) candidate region, and (iii) four strong peaks associated with leaf length (supported by P-values and QQ plot). QQ plots and more information on the lead SNPs are provided in Supplementary Material SM1.

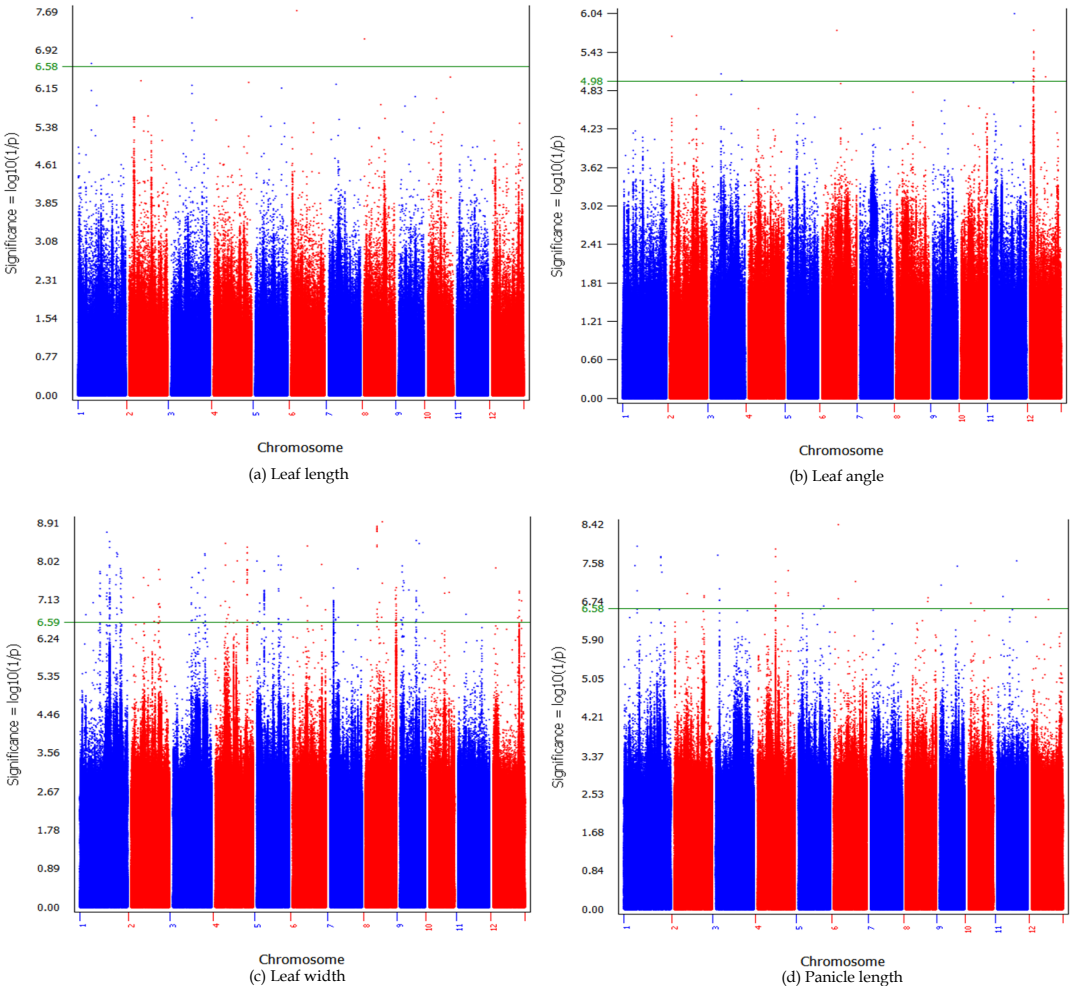


Figure 4: **GWAS on leaf and panicle traits.** We show the Manhattan plots for LL, LA, LW and PL.

4 Conclusions

Production of high-yield strains is crucial for meeting the continuously increasing food demand of the world population. Hybrid rice has been an effective way to meet this ever increasing food demand in Bangladesh. Despite a significant success in rice production, Bangladesh faces many challenges in the agricultural sector as it is becoming more densely populated day by day. In addition, climate change impacts like temperature rise, uncertain weather, prolonged dry season, irregular rainfall, frequent cyclones, sea-level rise,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

floods etc. are already being felt in Bangladesh. Therefore, understanding the genetic architecture underlying yield related traits as well as the impact of environmental factors are fundamental to the advancement of rice cultivars as the performance of various rice varieties vary with varying conditions of cultivation [31].

We presented genome-wide associated studies on 7 important yield-related traits using 183 rice varieties in Bangladesh. These GWASs are especially targeted for the Bangladeshi rice varieties, and thus it considers the interactions between genetic variations underlying yield-related traits and particular ecological environment of Bangladesh. In general, our GWAS results reveal some association signals for three grain shape related traits that were located close to the SNP loci that have been identified previously. Our results identified some chromosomes (e.g., chromosome 1, 6, 7, and 11) – to harbor lead SNPs – that are common for three grain shape related traits. We also identified SNP loci, which are strongly associated with both grain length and grain weight. This result could partially explain the genetic basis of correlation among the three traits (as demonstrated in Fig. 1), and provide useful information for genetic improvement of these traits by marker-assisted selection (MAS) [32–34]. As we discussed in our results section, there is discordance among the significant loci identified by various association studies performed on same traits. Differences in sample sizes and various types of rice varieties considered in different studies may be attributed to this disagreement. Another crucial factor is the ecological environment where the considered rice varieties were grown. Therefore, this study advances the state-of-the-art in rice research in Bangladesh. However, this study is limited in scope, and can be extended in various directions. We have leveraged the data from the 3K RGP project. Future studies need to collect rice materials, planted in various regions of Bangladesh under adverse ecological conditions to better elucidate the impact of specific environmental factors in genotype-phenotype association. Follow-up studies also need to investigate the candidate genes through functional genomics approach [1,35]. This study is limited to 7 yield-related traits. However, more information will be gained through GWAS of rice landraces as additional phenotypes are evaluated, especially the ones that are related to the adverse ecological environments of Bangladesh. To name a few, tolerance to prolonged flood, submergence, salinity, drought and cold are special features for various rice varieties in Bangladesh. As such, future studies need to sample a larger number of broadly representative varieties with special traits. For example, Rayada – a distinctive group of deepwater rice, totally endemic to certain area of Bangladesh and have multiple physiological features distinctly different from typical deepwater rice – could be potential resources of abiotic stress tolerance traits like flood, cold and drought [36–38]. Thus, we believe that this study will stimulate related future studies and will help identify beneficial genetic variations – which will enable the agricultural scientists to direct their efforts in developing elite varieties with desirable genetic compositions.

Supplementary Materials

Supplementary Material SM1: Additional results on leaf and panicle traits.
Data File SM2: Detailed information on the 183 rice varieties considered in this study.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

[1] Xuehui Huang, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, Chuanrang Zhu, Tingting Lu, Zhiwu Zhang, Meng Li, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, 42(11):961, 2010.

[2] Xuehui Huang, Yan Zhao, Canyang Li, Ahong Wang, Qiang Zhao, Wenjun Li, Yunli Guo, Liuwei Deng, Chuanrang Zhu, Danlin Fan, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, 44(1):32, 2012.

[3] Keyan Zhao, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiqul Islam, Andy Reynolds, Jason Mezey, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature Communications*, 2(1):1–10, 2011.

[4] Zhang Ya-fang, MA Yu-yin, Chen Zong-xiang, ZOU Jie, Chen Tian-xiao, LI Qian-qian, Pan Xue-biao, and Zuo Shi-min. Genome-wide association studies reveal new genetic targets for five panicle traits of international rice varieties. *Rice Science*, 22(5):217–226, 2015.

[5] Wanneng Yang, Zilong Guo, Chenglong Huang, Lingfeng Duan, Guoxing Chen, Ni Jiang, Wei Fang, Hui Feng, Weibo Xie, Xingming Lian, et al. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature Communications*, 5:5087, 2014.

[6] Filippo Biscarini, Paolo Cozzi, Laura Casella, Paolo Riccardi, Alessandra Vattari, Gabriele Orasen, Rosaria Perrini, Gianni Tacconi, Alessandro Tondelli, Chiara Biselli, et al. Genome-wide association study for traits related to plant and grain morphology, and root architecture in temperate rice accessions. *PloS One*, 11(5), 2016.

[7] Peng Zhang, Kaizhen Zhong, Zhengzheng Zhong, and Hanhua Tong. Genome-wide association study of important agronomic traits within a core collection of rice (Oryza sativa l.). *BMC Plant Biology*, 19(1):259, 2019.

[8] Xiao-ling LI, Yong-gen Lu, Jin-quan Li, Hai-ming Xu, and Muhammad Qasim Shahihd. Strategies on sample size determination and qualitative and quantitative traits integration to construct core collection of rice (oryza sativa). *Rice Science*, 18(1):46–55, 2011.

[9] LV Yang, Wang Yueying, Noushin Jahan, Hu Haitao, Chen Ping, Shang Lianguang, Lin Haiyan, Dong Guojun, Hu Jiang, Gao Zhenyu, et al. Genome-wide association analysis and allelic mining of grain shape-related traits in rice. *Rice Science*, 26(6):384–392, 2019.

[10] Xiaosong Ma, Fangjun Feng, Yu Zhang, Ibrahim Eid Elesawi, Kai Xu, Tianfei Li, Hanwei Mei, Hongyan Liu, Ningning Gao, Chunli Chen, et al. A novel rice grain size gene ossnb was identified by genome-wide association study in natural population. *PLoS Genetics*, 15(5):e1008191, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[11] Wanneng Yang, Zilong Guo, Chenglong Huang, Ke Wang, Ni Jiang, Hui Feng, Guoxing Chen, Qian Liu, and Lizhong Xiong. Genome-wide association study of rice (oryza sativa l.) leaf traits with a high-throughput leaf scorer. *Journal of Experimental Botany*, 66(18):5605–5615, 2015.

[12] 3000 Rice Genomes Project. The 3,000 rice genomes project. *GigaScience*, 3(1):2047–217X, 2014.

[13] Nickolai Alexandrov, Shuaishuai Tai, Wensheng Wang, Locedie Mansueto, Kevin Palis, Roven Rommel Fuentes, Victor Jun Ulat, Dmytro Chebotarov, Gengyun Zhang, Zhikang Li, et al. Snp-seek database of snps derived from 3000 rice genomes. *Nucleic Acids Research*, 43(D1):D1023–D1027, 2015.

[14] Locedie Mansueto, Roven Rommel Fuentes, Frances Nikki Borja, Jeffery Detras, Juan Miguel Abriol-Santos, Dmytro Chebotarov, Millicent Sanciangco, Kevin Palis, Dario Copetti, Alexandre Poliakov, et al. Rice snp-seek database update: new snps, indels, and queries. *Nucleic Acids Research*, 45(D1):D1075–D1081, 2017.

[15] Rongyu Huang, Liangrong Jiang, Jingsheng Zheng, Tiansheng Wang, Houcong Wang, Yumin Huang, and Zonglie Hong. Genetic bases of rice grain shape: so many genes, so little known. *Trends in Plant Science*, 18(4):218–226, 2013.

[16] Mayuko Ikeda, Kotaro Miura, Koichiro Aya, Hidemi Kitano, and Makoto Matsuoka. Genes offering the potential for designing yield-related traits in rice. *Current Opinion in Plant Biology*, 16(2):213–220, 2013.

[17] HX Huang and Q Qian. Progress in genetic research of rice grain shape and breeding achievements of long-grain shape and good quality japonica rice. *Chin J Rice Sci*, 31(6):665–672, 2017.

[18] Giang Thi Hoang, Pascal Gantet, Kien Huu Nguyen, Nhung Thi Phuong Phung, Loan Thi Ha, Tuan Thanh Nguyen, Michel Lebrun, Brigitte Courtois, and Xuan Hoi Pham. Genome-wide association mapping of leaf mass traits in a vietnamese rice landrace panel. *PloS One*, 14(7), 2019.

[19] Peng Wang, Guilin Zhou, Huihui Yu, and Sibin Yu. Fine mapping a major qtl for flag leaf size and yield-related traits in rice. *Theoretical and Applied Genetics*, 123(8):1319–1330, 2011.

[20] Hirokazu Tsukaya. Leaf shape: genetic controls and environmental factors. *International Journal of Developmental Biology*, 49(5-6):547–555, 2004.

[21] José Manuel Pérez-Pérez, David Esteve-Bruna, and José Luis Micol. Qtl analysis of leaf architecture. *Journal of Plant Research*, 123(1):15–23, 2010.

[22] Baoyan Jia, Xinhua Zhao, Yang Qin, Muhammad Irfan, Tae-heon Kim, Bolun Wang, Shu Wang, and Jae Keun Sohn. Quantitative trait loci mapping of panicle traits in rice. *Molecular Biology Research Communications*, 8(1):9, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[23] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[24] Steven N Goodman. Of p-values and bayes: a modest proposal. *Epidemiology*, 12(3):295–297, 2001.

[25] Shengwen Wang, Daniel Dvorkin, and Yang Da. Snpevg: a graphical tool for gwas graphing with mouse clicks. *BMC bioinformatics*, 13(1):319, 2012.

[26] Yoshihiro Kawahara, Melissa de la Bastide, John P Hamilton, Hiroyuki Kanamori, W Richard McCombie, Shu Ouyang, David C Schwartz, Tsuyoshi Tanaka, Jianzhong Wu, Shiguo Zhou, et al. Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6(1):4, 2013.

[27] Rakesh Singh, Ashok Kumar Singh, Tilak Raj Sharma, Aqbal Singh, and Nagendra K Singh. Fine mapping of grain length qtls on chromosomes 1 and 7 in basmati rice (oryza sativa l.). *Journal of Plant Biochemistry and Biotechnology*, 21(2):157–166, 2012.

[28] Kashif Aslam and Muhammad Arif. Ssr analysis of chromosomes 3 and 7 of rice (oryza staiva l.) associated with grain length. *Pak. J. Bot*, 46(4):1363–1372, 2014.

[29] Noriko Kinoshita, Masayuki Kato, Kei Koyasaki, Takuya Kawashima, Tsutomu Nishimura, Yuji Hirayama, Itsuro Takamure, Takashi Sato, and Kiyoaki Kato. Identification of quantitative trait loci for rice grain quality and yield-related traits in two closely related oryza sativa l. subsp. japonica cultivars grown near the northernmost limit for rice paddy cultivation. *Breeding Science*, page 16155, 2017.

[30] Xiaoqiong Li, Yu Wei, Jun Li, Fangwen Yang, Ying Chen, Yinhua Chen, Sibin Guo, and Aihua Sha. Identification of qtl tgw12 responsible for grain weight in rice based on recombinant inbred line population crossed by wild rice (oryza minuta) introgression line k1561 and indica rice g1025. *BMC Genetics*, 21(1):10, 2020.

[31] S Sreedhar, T Dayakar Reddy, MS Ramesha, et al. Genotype x environment interaction and stability for yield and its components in hybrid rice cultivars (oryza sativa l.). *International Journal of Plant Breeding and Genetics*, 5(3):194–208, 2011.

[32] KA Steele, Adam Huw Price, HE Shashidhar, and JR Witcombe. Marker-assisted selection to introgress rice qtls controlling root traits into an indian upland rice variety. *Theoretical and Applied Genetics*, 112(2):208–221, 2006.

[33] KK Jena and DJ Mackill. Molecular markers and their use in marker-assisted selection in rice. *Crop Science*, 48(4):1266–1276, 2008.

[34] Zhe Zhang, Ulrike Ober, Malena Erbe, Hao Zhang, Ning Gao, Jinlong He, Jiaqi Li, and Henner Simianer. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PloS One*, 9(3), 2014.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[35] Qifa Zhang, Jiayang Li, Yongbiao Xue, Bin Han, and Xing Wang Deng. Rice 2020: a call for an international coordinated effort in rice functional genomics. *Molecular Plant*, 1(5):715–719, 2008.

[36] ANM Rubaiyath Bin Rahman and Jianhua Zhang. Rayada specialty: the forgotten resource of elite features of rice. *Rice*, 6(1):41, 2013.

[37] Antonio T Perez and Muhammad Nasiruddin. Field notes on the rayadas: a flood-tolerant deepwater rice of bangladesh. In *Proceedings of the International Seminar on Deepwater Rice*, number 15, pages 87–91, 1974.

[38] Jean-Christophe Glaszmann. Isozymes and classification of asian rice varieties. *Theoretical and Applied Genetics*, 74(1):21–30, 1987.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Material to Genome Wide Association Studies on 7 Yield-related Traits of 183 Rice Varieties in Bangladesh

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

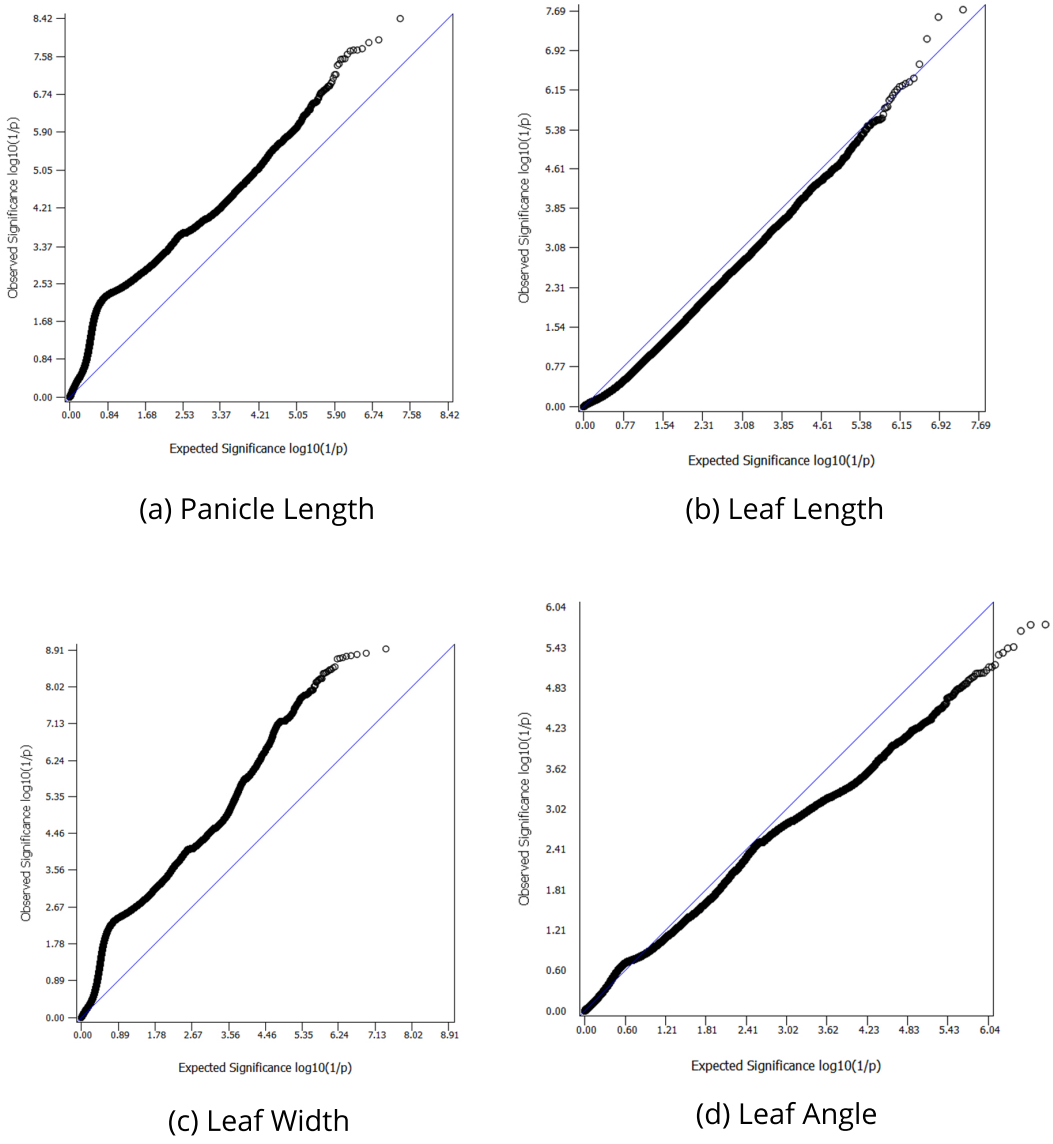


Figure S1: **QQ Plots on leaf and panicle traits.** The horizontal axis shows the $-\log10$-transformed expected P-values, and the vertical axis indicates $-\log10$-transformed observed P-values

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.22.393074; this version posted November 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table S1: **Genome-wide lead SNPs significantly associated with panicle length and leaf related traits.** For each trait, we show the significantly associated SNPs. We also show the candidate genes, harboring these significant SNPs. Beta denotes the coefficient from a fit. For LW, we show the top 20 lead SNPs (among a total of 94 significant SNPs).

Trait	Chr	SNP ID	SNP (bp)	Minor Allele	Major Allele	MAF	P	Beta (β)	T
LA	11	336082971	19390414	A	T	0.28	9.20E-07	-2.41	-5.4
	12	349301237	3587574	C	A	0.14	1.68E-06	2.01	5.22
	6	194632495	13549375	G	A	0.12	1.69E-06	4.05	5.26
	2	45302658	2031735	G	A	0.43	2.09E-06	-1.31	-5.13
	12	349364246	3650583	A	G	0.07	4.43E-06	2.37	4.97
	12	348976907	3263244	A	T	0.09	7.22E-06	2.17	4.86
	3	93129009	13920836	A	G	0.28	8.12E-06	1.27	4.78
	12	349373785	3660122	A	G	0.19	9.00E-06	1.29	4.75
	12	359971828	14258165	C	T	0.05	9.01E-06	3.31	4.82
	3	112939373	33731200	A	T	0.24	1.04E-05	2.72	4.73
LL	6	186252827	5169707	C	T	0.22	2.04E-08	-0.60	-6.00
	3	98618701	19410528	C	T	0.21	2.84E-08	-0.64	-5.95
	8	242276041	246513	T	C	0.11	7.52E-08	-0.81	-5.71
	1	12035025	12035025	G	A	0.17	2.32E-07	0.61	5.47
LW	8	257488768	15459240	T	C	0.06	1.23E-09	4.96	6.51
	8	252585885	10556357	A	G	0.05	1.56E-09	0.30	6.50
	1	24422813	24422813	A	T	0.06	2.15E-09	5.25	6.38
	9	285199478	14726928	C	T	0.07	3.34E-09	5.18	6.29
	1	26927183	26927183	G	A	0.08	3.53E-09	3.66	6.28
	4	125098182	9476190	C	T	0.05	3.89E-09	3.82	6.29
	9	287877924	17405374	G	A	0.06	3.92E-09	5.16	6.25
	6	194707456	13624336	T	C	0.07	4.46E-09	3.43	6.26
	4	145570760	29948768	T	C	0.09	4.77E-09	2.31	6.21
	1	33379573	33379573	C	T	0.06	6.43E-09	5.23	6.18
	3	109144323	29936150	T	C	0.06	6.76E-09	3.56	6.14
	1	34209824	34209824	A	G	0.12	6.93E-09	0.20	6.19
	5	171146372	20021686	C	T	0.07	7.70E-09	4.38	6.11
	1	33702095	33702095	A	G	0.06	7.86E-09	5.17	6.11
	4	136194663	20572671	T	A	0.06	9.89E-09	5.07	6.07
	5	151306979	182293	G	A	0.07	9.99E-09	4.67	6.06
	6	207956689	26873569	T	C	0.06	1.20E-08	4.95	6.03
	5	173727106	22602420	A	C	0.06	1.24E-08	5.01	6.02
	9	272571916	2099366	A	G	0.07	1.29E-08	4.96	6.03
	12	347991169	2277506	A	G	0.06	1.43E-08	5.01	6.00
PL	6	185302850	4219730	A	G	0.13	3.78E-09	24.31	6.29
	1	12092594	12092594	G	A	0.29	1.12E-08	2.52	6.04
	4	132814753	17192761	C	T	0.14	1.30E-08	-2.10	-6.06
	3	81204448	1996275	T	G	0.07	1.77E-08	-4.32	-6.03
	1	34026055	34026055	G	T	0.50	1.90E-08	22.45	5.94
	11	335095423	18402866	C	T	0.10	2.36E-08	3.31	5.91
	1	10059575	10059575	T	C	0.13	3.00E-08	-2.55	-5.92
	9	286628275	16155725	A	G	0.10	3.07E-08	-3.77	-5.89
	4	144826279	29204287	C	T	0.24	3.84E-08	1.86	5.79
	1	34989124	34989124	C	T	0.10	4.16E-08	3.34	5.83
	6	201474144	20391024	T	C	0.06	6.65E-08	4.26	5.73
	9	271855609	1383059	G	A	0.18	8.00E-08	20.32	5.64
	3	82872615	3664442	A	G	0.09	9.65E-08	-3.11	-5.63
	2	55441587	12170664	T	C	0.05	1.23E-07	39.04	5.55
	2	71500571	28229648	T	C	0.33	1.36E-07	1.80	5.56
	11	322211290	5518733	C	A	0.38	1.42E-07	2.98	5.57
	8	262899554	20870026	T	C	0.07	1.51E-07	3.52	5.57
	12	360565118	14851455	T	C	0.10	1.65E-07	-3.66	-5.54
	3	82894142	3685969	T	G	0.22	1.71E-07	-1.59	-5.48
	8	262572164	20542636	T	C	0.33	1.83E-07	2.64	5.48
	10	295893660	2408390	A	G	0.27	1.98E-07	2.42	5.45
	5	174896782	23772096	A	G	0.08	2.30E-07	3.53	5.46