Combination of favorable gene haplotypes plays an important role in rice hybrid grain quality traits based on genome-wide association analysis

Lanzhi Li ^{1†}, Xingfei Zheng ^{2,3†}, Xueli Zhang ^{1†}, Kai Xu ¹, Shufeng Song ⁴, Jing Su ¹, Chao Wu ¹, Zirong Li ¹, Wenbang Tang ⁵, Ying Diao ², Zhonghai Tang ⁶, Zheming Yuan ^{*1}, Zhongli Hu^{*2},

¹ Hunan Engineering & Technology Research Center for Agricultural Big Data Analysis & Decision-making, Hunan Agricultural University, Changsha, 410128; ² State Key Laboratory of Hybrid Rice, Wuhan University, Wuhan, 430072; ³ Hubei Key Laboratory of Food Crop Germplasm and Genetic Improvement, Food Crop Institute, Hubei Academy of Agricultural Sciences, Wuhan 430064, China; ⁴ State Key Laboratory of Hybrid Rice, Hunan Hybrid Rice Research Center, Changsha 410125, China; ⁵ Rice Research Institute, Hunan Agricultural University, Changsha, Hunan, China; ⁶ College of Food Science and Technology, Hunan Agricultural University, Changsha 410128, China

[†] co-first author

^{*} co-corresponding author

Email: huzhongli@whu.edu.cn

Running head: Genetic dissection of rice grain quality traits based on genome-wide association study

Summary:

Yield level and grain quality determine the commercial potential of rice (*Oryza sativa* L.)

varieties. Mining and using genes that control important rice grain quality characteristics are major tasks for plant breeders. Here, a genome-wide association study was conducted to determine the genetic bases of 12 rice grain quality traits in 113 varieties and 565 testcross hybrids. A total of 56 significant SNPs were associated with 9 of the traits in variety phenotypic, general combining ability, testcross hybrid phenotypic and mid-parental heterotic datasets, from which 45 novel loci were identified. The cumulative effects of superior alleles or favorable haplotypes of genes closest to significant quality trait-associated loci were found in the four datasets. Additionally, the favorable gene haplotypes performed better than those of superior alleles in the variety and testcross hybrid datasets. Pyramiding the favorable haplotypes of five cloned rice grain quality genes resulted in a very low amylose content and high yield in the latter. These testcross hybrids had rice grain qualities similar to their parental lines but with much higher yields. The amylose content, grain width and grain length predictions for potential hybrids among the 113 varieties using genomic selection based on the BayesB method revealed a performance trend similar to those the testcross hybrids in our study. Thus, the selection of combination of favorable rice quality-related gene haplotypes is recommended to breed hybrids with high yields and elite grain qualities.

Keywords: rice grain quality; genome-wide association study; combining ability; heterosis;

hybrids

Introduction

More than half of the world's population consume rice as their staple food. With improvements in living standards, people are paying more attention to food quality and a healthy lifestyle, which requires rice (*Oryza sativa* L.) to have a higher end-use cooking quality

(Huggins et al., 2018). Grain quality is a complex trait regulated by interacting genetic and environmental factors. Mining and utilizing genes that control important quality characteristics of rice are major tasks of plant breeders (Huang et al., 2010; McCouch et al., 2010; Tung et al., 2010). However, the simultaneous improvement of yield and end-use quality remains a challenge for rice breeders because yield and quality are typically negatively correlated (Huggins et al., 2018). In general, the yields of hybrid rice are greater than those of conventional rice varieties, but the grain quality is not improved. To obtain both high-quality and high-yield rice varieties, the quality traits of hybrid rice need to be genetically dissected.

A high milled brown rice rate, translucent grains and good taste are desired on essentially all markets, and these grain characteristics are the bases of evaluation indexes for cooking rice. Rice grain quality has been classified into milling, appearance, cooking and nutritional categories (Fitzgerald and Resurreccion, 2009; Huggins et al., 2018). In recent years, with the rapid development of molecular biology-based theories and technology, many rice grain quality trait-related genes have been cloned and localized, and their mechanisms and functions have also been investigated. Having genomic markers linked with various grain quality traits in rice will help breeders combine crop productivity with desirable end-use quality traits targeted for specific markets (Huggins et al., 2018). With molecular marker-assisted selection, breeders can introduce good quality genes into high yield varieties and greatly improve the breeding efficiency.

Grain size is closely related to the yield (Che et al., 2016; Asante, 2017). At present, dozens of grain size-related genes have been isolated from different rice germplasm resources, such as the genes *GS3*, *GL3.1* and *GW7/GL7*, controlling grain length (GL; Mao et al., 2010; Wang et

al., 2015a; Wang et al., 2015b), the genes *GW2*, *GW5/qSW5* and *GS5*, controlling grain width (GW; Song et al., 2007; Wang et al., 2008; Li et al., 2011), and the genes *GS6*, *GS9*, *TGW6* and *GW8/SPL16*, controlling grain size (Sun et al., 2013; Ishimaru et al., 2013; Song et al., 2015; Che et al., 2016). Chalky grains are considered as low quality because of their poor appearance and the undesirable cooking and milling qualities (Zhao et al., 2018). The rice gene *OsRab5a* regulates endomembrane organization and storage protein trafficking in rice endosperm cells, which affects the formation of amyloplasts (Wang et al., 2012); and *Chalk5* encodes a vacuolar H⁺-translocating pyrophosphatase, which influences grain chalkiness in rice (Li et al., 2014).

The amylose content (AC) has the greatest influence on the cooking and consumption qualities of rice. The synthesis of rice amylose is catalyzed by granule-bound starch synthase protein that is encoded by the waxy gene *Wx* (Smith et al., 1997). There are several alleles of the waxy gene, including *Wx*, *Wx^a*, *Wx^b*, *Wx^{mp}*, *Wx^{op}*, *Wxⁱⁿ* and *Wx^{mq}*. Zhou et al. (2016) cloned a practical resistant starch gene based on the physical map. Rice varieties with an intermediate gel temperature, which is predominantly determined by amylopectin structure, are generally preferred by consumers. The gene, starch synthase II is the major determinant of gel temperature (Asante, 2017).

Protein content (PC) is a major index of rice grain nutritional quality. Generally, the PCs in *indica* rice varieties are higher than those in *japonica*. In total, 19 QTLs have been linked to rice grain PC. Liao et al. (2019) found that over-expressed *OsAPP6* could facilitate the roots absorption of NH₄⁺, which improved rice grain protein synthesis.

QTL mapping and cloning provide important information for dissecting main-effect genetic loci, but it is often difficult to mine the minor-effect genes associated with traits. In

addition, only the genetic information of the two parents are included in a QTL analysis, making it difficult to identify the network relationships among complex genes in narrow germplasm resources (Huang et al., 2018, Fang et al., 2018). A genome-wide association study (GWAS) has a higher resolution than QTL mapping because it includes the recombination information accumulated by natural populations during long-term evolution. Fine positioning of the QTL can be achieved in an association analysis, even directly targeting the gene itself (Huang et al., 2018). GWASs have been widely applied in crops, such as wheat, corn, potato, adzuki bean, perennial ryegrass, rice and mustard, to decipher the genetic bases of quantitative traits (Huang et al., 2010; Xie et al., 2010; Huang et al., 2015; Huang et al., 2016; Huang et al., 2018), Zhao et al. (2010) used a chip containing more than 36,000 high-quality SNPs to perform a GWAS on 34 traits in more than 400 rice varieties in 82 countries. Some known genes were found, such as *qSW5* and *GS3*, which control grain size. With the development of high-throughput sequencing technology, GWASs based on whole-genome sequencing have been widely used in research on rice agronomic and quality traits. In 2015, Huang et al. directly sequenced 1,495 hybrid rice varieties in China's main rice producing areas and used 38 agronomic traits, such as yield, quality and disease resistance, of these hybrid rice varieties. In 2018, Misra et al. conducted a GWAS on cooked rice's quality traits (adhesion, hardness, elasticity, viscosity and AC) for 236 *indica* rice varieties from 37 countries. A total of 224 significantly associated SNP loci were detected, of which 97 were validated by at least two GWAS methods. GWAS of three rice diversity panels were performed by Huggins et al. in 2018. Their results revealed large- and small-effect loci associated with known genes and previously uncharacterized genomic regions and suggested that multiple grain quality traits

were inherited together.

At present, GWAS research on rice quality traits is mainly conducted using conventional rice accessions, while the study of hybrid rice quality is limited. To obtain both high-quality and high-yield rice varieties, it is necessary to decipher the genetic bases of hybrid rice grain quality. In this study, a GWAS analysis was performed on the grain quality traits of 113 *indica* rice varieties in China and their 565 testcross hybrids based on the North Carolina II (NCII) design. The aim of this study is to provide greater insights into effective breeding strategies for the grain quality improvement of rice hybrids.

Results

Rice grain quality trait phenotypes and their relationships

Phenotypic distributions of the 12 quality traits in testcross hybrids and their male parental lines are shown in Fig. 1a and Supplementary Figure 1 and 2. The 12 traits in the four datasets (V, TC, Gca and Hmp) were distributed continuously with diverse variation levels, indicating that complex genetic mechanisms underlie rice grain quality traits. Rice grains with low CD and PGWC, and high TD, PC and ASV, are desired for elite quality rice. In this study, the mean values of CD and PGWC in the testcross hybrids were higher than those of their male parental lines, but the former had lower TDs, PCs and ASVs. This result partially confirmed that hybrids have poor rice grain qualities when compared with those of conventional varieties. For the remaining seven traits, testcross hybrids showed similar performances with those of their male parental parents.

Correlation analyses for every pair of traits were performed within the four datasets. The

trends of most traits' relationships were consistent within the V and Gca datasets (Fig. 1b), and within the TC and Hmp datasets (Supplementary Figure 3). Additionally the significance levels of the correlations varied among traits in the same dataset. GL and GW are closely related to yield, and these two traits are always negative correlated. In this study, GL was significantly negatively correlated with CD, AC and TD in all the datasets, while GW was positive correlated with them. Generally, grains with a low CD, long GL and low AC are desired in rice. The correlation coefficient analysis results confirmed the trend and indicated that the germplasm materials in this study may have been preserved by long artificial selection. We found that GW and GL were not significantly associated with ASV, BRR, MRR or HRR. Thus, breeders should consider the relationships among traits when conducting rice breeding for high yield and grain quality.

Structure of the NCII population

The NJ tree suggested that there were population differences within the male parental lines (Supplementary Figure 4); however, no clear clades were shown, which indicated that these materials have a complicated breeding history. The male sterile lines and restorer lines were not separated into two divergent clusters, which further indicated the wide variation within our materials. Five branches of testcross hybrids were detected at $K = 5$ (Supplementary Figure 5). The testcross hybrids derived from the same male parent line clustered in the same clades in the NJ tree, which was similar to the findings of Huang et al. (2015) and Zhen et al. (2017). Therefore, when performing a GWAS analysis of TC and Hmp, special attention should be paid to the influence of population structure. The PCA of the 120 parental lines revealed a wide

variation among them (Supplementary Figure 6), while a PCA revealed a highly-structured genetic relationships among the TC lines (Supplementary Figure 7). The first and second genomic PCs mainly separated the TC lines into two groups, while the third and fourth PCs separated them into four small groups consistent with their female parental origins, except for 3A and Y58S, which were together. To correct for subpopulation structure, we integrated the first three PCs as covariates within the model and performed a GWAS across all the accessions in the V, Gca, TC and Hmp datasets.

Whole-genome screening for significantly associated loci

To reduce false positive or negative results, four algorithms: general linear model, mixed linear model, compressed mixed linear model-based EMMAX (Kang et al., 2010) and fixed and random model circulating probability unification model in software GAPIT were performed in the GWAS pipeline. Then, the Q-Q (quantile-quantile) plot and the Manhattan graphs obtained by each model (Figs 2 and 3, Supplementary Figures. 8–31) were compared to determine the best results for each trait in all the datasets ($12 \times 4 = 48$ datasets in total), which were used for subsequent analyses (Supplementary Table2). Using a strict Bonferroni correction, the threshold was set to $p < 2.64 \times 10^{-8}$ ($0.01/\text{total number of SNPs}$), and 56 significant SNPs were detected, involving 9 traits and 24 datasets (Table 1).

In our research, four of eight significant markers detected for GL harbored or were located near reported genes (*GS3*, *GL3.2*, *GL7*, *PGL1* and *PGL2*) associated with GL (Fig. 2). For GW, SNP chr05_5359838 was simultaneously detected as being significant in Gca and TC. It was located near the GW main-effect gene *GW5*. SNP chr03_35360872 was harbored in gene *qTGW3*, which regulates rice GW (Table 1). The waxy gene is involved in amylose synthesis

in the endosperm and is located on chromosome 6. Approximately 34 kb apart from it, the significant SNP chr06_1731808 was identified for AC in the TC dataset (Fig. 2). For chalkiness, SNP chr05_4591773 for CD was detected as being significant in the Gca and TC datasets. It was approximately 120 kb from the *chalk* gene, which is the main-effect gene for grain chalkiness regulation. SNP chr05_4140727 for PGWC in the Gca dataset was 80 kb from *chalk*. The other 49 significant markers for grain quality traits were first detected here.

Assessing pleiotropy between traits

Pleiotropy and LD play important roles in identifying correlations among phenotypes. In this study, we found that several SNPs or their nearby regions associated with different traits (Supplementary Table 3), which helped us dissect the genetic architecture of the correlations across different traits. For example, the chr03_16720836 locus was simultaneously detected to be correlated with GL in Gca, GL in TC and GLWR in Gca, and 0.32 kb from this locus, SNP chr03_16721158 was detected to be correlated with GLWR in TC. These two loci are in the vicinity of the cloned gene *GS3* (position: chr03:16729501 to 16735109), which is the major gene controlling rice GL and GW, and also a minor-effect QTL controlling rice GW and grain filling. In addition, the same traits in different datasets had overlapping peaks within a given genomic region. SNP chr06_1621853 on chromosome 6 associated with AC in Gca, which is located at a close distance from SNP chr06_1629652, which was significantly associated with AC in the V dataset. *OsACS6*, which regulates the size and morphology of starch granules in endosperm cells, is located near these two sites. These results are not only consistent with other

researcher's findings, but also suggest that some traits share a genetic basis that allows a general combining ability or heterosis. Some quality traits, or the same trait's general combining ability and heterosis, are more likely to be co-inherited, and these important loci may be used in breeding to improving hybrid rice quality.

Dominance degree

To dissect the genetic basis of rice grain quality trait heterosis, we calculated the dominance degree of the top 100 SNPs of the GWAS results for all the data using descending $-\log(P \text{ values})$. Generally, the dominance degree varied in the four datasets. Except for PGWC and PC, all the traits showed A followed by PD in the V dataset (Supplementary Figure 32). In the Gca dataset, on the whole, A and PD accounted a greater proportion than D and OD. A accounted for a large proportion of the traits GW, ASV, MRR and HRR, while PD mainly accounted for GL, PGWC, CD and BRR. Similar trends in the dominance degrees of quality traits occurred in the Gca dataset. The four types of dominance existed and varied among traits, but in general, A and PD were still relatively large in the TC dataset. The OD proportion increased in the Hmp dataset when compared with TC dataset. Compared with the phenotypic values of the testcross hybrids, the Hmp value is more suitable for detecting the heterozygotic genetic basis. Because only dominant effect involved in Hmp at single locus level, while additive and dominant effect in testcross hybrid phenotypic performance (Li et al., 2008). Thus, A is main genetic basis of rice quality traits in these accessions (male parental lines in this study), while OD mainly underlies the heterotic genetic basis.

Frequency of superior alleles of associated loci and favorable haplotype of gene

The frequency of superior alleles of associated loci in V and TC datasets was investigated

following the method of Huang et al. (2015), in which the allele with the better quality performance (for example, greater GL, greater GW or lower AC) was defined as the superior allele. Because of the low LD rate of decay in rice, a GWAS cannot resolve a single gene (Huang et al., 2010), we investigated the frequency of the favorable haplotype of the closest gene to the associated loci. Similarly, the haplotype having the better quality performance was defined as the favorable haplotype of the gene. Our results (Fig. 4) showed that, overall, the performance of paternal lines or testcross hybrids of each trait improved as the number of superior alleles or favorable haplotypes increased. Additionally, the performance of the favorable haplotype of the gene outperformed the superior allele of the SNP (lower AC, similar GW and longer GL) in the V dataset. The TC lines in which favorable haplotype of genes were pyramided also showed good rice grain quality that approached those of the male parental lines but much higher yields, which suggested that breeders may select hybrids having high yields and elite quality.

Indirect association of quality traits owing to a high LD level

In this study, we calculated the LD block using a pairwise comparison of SNP markers separated by 5 Mb. The average LD ($r^2 < 0.5$) in the parental lines was ~200 kb (Supplementary Figure 33). The low rate of LD decay in rice means that a GWAS cannot resolve a single gene (Huang et al., 2010). Several significant SNPs were identified in the vicinity of *GW5*, *GS3*, *Wx*, *OsACS6* and *ALK*. Among them, *Wx*, *OsACS6* and *ALK* influence the ACs of rice grains (Supplementary Figure 34–S36), and *GW5* and *GS3* are closely related to rice grain yield. There

are various haplotypes in the V and TC datasets, which may explain why the SNP markers could not achieve statistical significance in those regions (Ye et al., 2018). We ignored other varieties or testcross hybrids containing minor haplotypes. For *Wx* (Fig. 5), Haplotypes A, B and C, which included most varieties in the V dataset, corresponded to haplotypes J, K and L, respectively, in the TC dataset. The varieties having haplotype C had lower ACs (18.50%) and yields (26.15 g) than those having haplotype A or B, while haplotype L accessions had low ACs (18.95%) but also high yields (41.25 g) (Fig. 5a and b). Thus, haplotype L was determined to be optimum for *Wx* in the TC dataset.

A similar phenomenon was observed in the coding region of genes *OsACS6* and *ALK*. For *OsACS6*, haplotype M accessions were the optimum (with low ACs and high yields) in the TC dataset, which corresponded to the haplotype B in the V dataset (Supplementary Figure 37). It should be noted that the varieties having haplotype B in the V dataset had low ACs (18.90%) and relatively low yields (23.42 g), but corresponded to haplotype M of the testcross hybrids in the TC dataset, which had low ACs (19.09%) but high yields (37.33 g). Thus, the parental lines with poor performances can still generate elite hybrids owing to heterosis.

For *ALK*, accessions of haplotype DD were optimum in the TC population, corresponding to haplotype D in the V dataset (Supplementary Figure 38). The accessions simultaneously having three superior haplotypes (haplotypes L, M and D for *Wx*, *OsACS6* and *ALK*, respectively) outperformed (with 18.29% AC and 39.34 g yield) other accessions having only one elite haplotype in the TC dataset.

For *GW5*, accessions of haplotype H were optimum in the TC population, corresponding to haplotype H in the V dataset (Supplementary Figure 39). For *GS3*, accessions having haplotype

J were optimum in the TC population, corresponding to haplotype C in the V dataset (Supplementary Figure 40).

Predicting untested crosses using BayesB

To investigate the quality trait performance of potential hybrids among the 113 inbred lines in our study, we conducted genomic selection predictions using the BayesB method. All the potential hybrids' predicted values are shown in Supplementary Tables 5–7. The average predicted AC and yield value of the 224 hybrids with accumulating favorable haplotypes of genes (*Wx*, *OsACS6* and *ALK*) were 17.3% and 42.52 g, which has similar with results from testcross hybrids in this study. When combined them with favorable haplotypes of genes (*GW5* and *GS3*), 18 potential hybrids showed average predicted AC and yield values of 17.06% and 43.16 g. Thus, accumulating favorable haplotypes of important genes related to rice quality and yield traits may help us produce desirable rice hybrids.

Discussion

In this study, the GWAS method was used to analyze the genetic basis of 12 rice grain quality traits in 113 parental lines and their 565 testcross hybrids. A total of 56 significant SNPs were associated with 9 of the 12 traits in the V, Gca, TC and Hmp datasets. Among them, three significant SNPs were located within or near the cloned quality genes, while the potential roles of the other SNPs in affecting rice grain quality were newly discovered. They provide new information for rice quality trait breeding. A few of these associated loci were detected simultaneously in multiple datasets of the same trait, but most of the significant loci in the

different datasets were unique. This indicated that the quality traits' general combining ability and heterosis are quantitative traits, controlled by multiple genes; however, the phenotypic performance of inbred breeding lines and their general combining ability, as well as phenotypic performance of hybrids and their heterosis, have different genetic bases for rice grain quality. These findings are consisted with the QTL mapping and GWAS analyses of rice agronomic traits presented in previous research using the NCII (north Carolina design II) and TTC (triple test crosses) designs (Li et al. 2008, Qu et al., 2012, Chen et al., 2019).

The dominance degrees of the top 100 associated loci suggested that the phenotypes of quality traits in inbred lines were mainly controlled by A, while the dominance degree distribution trends of grain quality traits of parental lines and testcross hybrids' phenotypic performances varied in traits, but were mainly controlled by A and PD on the whole. More PD loci existed than other types of loci for GL, GW, GLWR, CD and BRR, while more A loci existed than other type of loci for PGWC, AC, ASV, MRR and HRR. The heteroses of GW, CD, TD and AC (measured by the mid-parental values in this study), GL, PC and BRR, and of GLWR, ASV, MRR and HRR are mainly controlled by OD, PD and A, respectively, which is different from the genetic basis of hybrid rice' agronomic traits studied by Huang et al (2015). They found fewer OD loci than those with incomplete dominance, which implies that dominance is an important contributor to the heterosis of rice grain quality in hybrid breeding.

The average mean values of quality traits increased along with the number of superior alleles of significant associated loci in the four datasets. Similarly, Huang et al. (2015) studied the superior alleles of top 100 GWAS loci (order them in descending of $-\log(P \text{ value})$) associated with agronomic traits of hybrid progenies and parents. They suggested that the

heterosis of hybrid progeny might be superior to that of inbred rice varieties for superior allele accumulation. Our study shows that in terms of quality traits, the phenotypic performance and Gca of inbred rice varieties (in this study, repented as male parental lines), as well as the phenotypic performance and the heterosis of hybrid progeny, may be the result of the accumulation of rare superior alleles. Chen et al. (2019) also found that high Gca parental lines or high Sca (special combining ability) combinations are caused by pyramiding favorable alleles.

Although significant rice grain quality trait-associated loci were detected, pyramiding associated loci of superior alleles does not guarantee improved rice quality. Because of LD, the genes may only be indirectly associated with targeted traits. However, the GWAS results need further analyses, such as transcript and genetic transformational analyses. In this study, the combinations of favorable haplotypes of three indirectly associated genes of AC (*Wx*, *ALK* and *ACS6*), showed the desired ACs in the male parental lines (18.62%), and surprisingly, the lower AC (15.97%) and higher yield were found in testcross hybrids with the favorable haplotype. Although varieties with combinations of superior alleles of the four significant SNPs associated with AC in the V dataset had low ACs (15.8%), their corresponding lines in the TC dataset had increased ACs, at 17.92% on average. These procedures were also conducted for the traits GL and GW (Fig. 4), and similar trends were observed. Generally, the parental lines with favorable combinations of gene haplotypes (or a favorable haplotype of an individual gene) controlling grain traits crossed with male sterile lines generated hybrids with similar grain quality performances as their male parent but higher yield performances. Thus, hybrids with an improved end-use quality may result from pyramiding favorable gene haplotypes that influence

rice quality traits.

Genomic selection predictions of AC, GL, GW and yield for all 6,328 potential hybrids between the 113 inbred lines were conducted. Among the potential hybrids, selecting those with favorable haplotypes of *Wx*, *ALK*, *ACS6*, *GW5* and *GS3* lead to elite ACs (17.06%) and high yields (43.16 g) on average. We are planning to generate these 18 potential hybrids in a future trial experiment. We hope our research may facilitate dissecting the genetic bases of grain quality traits of rice hybrids and help breeders select and produce ideal hybrids.

Experimental procedures

Materials and field planting

In accordance with the NCII genetic mating design, 115 *indica* rice accessions, as male parents, were crossed with five male sterile lines, as female parents, to produce testcross hybrids. In 2013, male parental lines (Supplementary Table 1) and testcross hybrids were planted in Wuhan, China. The field trials were designed as randomized blocks, repeated twice, with 20 plants per field planted at a density of 5×8 inches. Field management, including irrigation, fertilizer application and pest control, followed normal agricultural practices. The grains were harvested when fully ripe. In total, 113 parental lines (without samples V2 and V3 in Supplementary Table1) and 565 (113×5) testcross hybrids were obtained for phenotyping.

Phenotyping

After the materials matured, three plants with uniform growth were selected from the middle eight plants, dried, and stored at room temperature for 3 months. They were then sent to the Agricultural College of Hunan Agricultural University for rice grain quality evaluation.

In accordance with the method of Lou et al. (2009), 12 rice quality traits including the grain length (GL, mm), grain width (GW, mm), GL/GW ratio (GLWR), chalkiness degree (CD, %), percentage of grain with chalkiness (PGWC, %), transparency (transparent degree, TD), amylose content (AC %), alkali spreading value (ASV), PC (total protein content %), brown rice ratio (BRR, %), milled rice ratio (MRR, %) and whole milled rice rate (head rice ratio, HRR, %) of male parental lines and testcross hybrids were determined. The 12 quality traits of female parental lines were measured with those of the five male sterile lines' seeds. The yield production per plant of male parental lines and testcross hybrids were also evaluated.

Resequencing, genotyping and imputation

For each parental line, the genomic DNA was prepared from a single plant for sequencing. A sequencing library was established in accordance with the Illumina protocol. The genomes of 118 parental lines were sequenced on an Illumina HiSeq2500 platform with 11× genome coverage on average. 'Nipponbare' was used as the reference genome, and BWA software (Li and Durbin, 2009) was used for all paired-end read mapping. Then SNP calling and quality control were conducted by deleting SNPs with minor allele frequency > 5% and missing rate > 20%. In total, 1,894,012 SNPs were obtained and analyzed by NPUTE (ver 4.0, Roberts et al., 2007) with no missing genotype. Testcross hybrid genotypes were inferred using parental SNP genotypic information.

Phylogenetic and population structures

Neighbor-joining (NJ) trees and principal component analysis (PCA) plots were used to infer the structures of the 118 rice parental lines or 565 testcross hybrids. A pairwise distance matrix derived from the simple matching distance for all the SNP sites was calculated to construct

unweighted NJ trees using the software SNPhylo (version 20140116, Lee et al., 2014), and phylogenetic trees were drawn by iTOL online (<http://itol.embl.de/>). The program Admixture (version 1.3) was used to calculate varying levels of K ($K = 1-10$; Alexander et al., 2009). The PCA was conducted using GAPIT (version 2.0, Tang et al., 2016) and was separately calculated using SNPs present in parental lines and testcross hybrids. Genome-wide linkage disequilibrium (LD) was estimated using pairwise r^2 values between SNPs, which were calculated using the `--r2 --ld-window99999 --ld-window-r2 0` command in PLINK (version 1.9; Purcell et al., 2007).

Phenotypic and heterotic statistics and a GWAS analysis

Histograms, boxplots and correlations were constructed using grand means for V or TC. Heterosis was evaluated in testcross hybrids using Hmp, as follows:

$$Hmp_{ij} = Y_{ij} - (V_i + S_j)/2,$$

where Y_{ij} represents the phenotypic value of the testcross hybrid derived from the i th male parent and j th female parent, V_i represents the phenotype of the i th male parental lines, and S_j represents the j th sterile line phenotypic value. The Gca was measured using the Gca effect (g_i), as follows:

$$g_i = \bar{y}_i - \bar{y} \cdot \cdot,$$

where g_i represents the i th maternal parent's Gca effect, $\bar{y}_i$ represents the mean phenotypic value of testcross hybrids that derived from the i th male parent, and $\bar{y} \cdot \cdot$ represents the mean phenotypic value of all the testcross hybrids. A GWAS analysis was separately conducted on each trait in the four datasets V, Gca, TC and Hmp using the compressed mixed linear model with the GAPIT software (version 2.0; Zhang et al., 2010), general and mixed linear models in MVP (<https://github.com/XiaoleiLiuBio/rMVP>) and FarmCPU (fixed and random model

circulating probability unification)(Liu et al., 2016). In addition to the specific parameters of these four algorithms, the PCA was set to 3 for uniformity, and we defined a whole-genome significance cutoff using the Bonferroni test threshold of $P < 2.64 \times 10^{-8}$ [$-\log_{10}(P \text{ value}) = 7.58$, the Bonferroni correction for multiple tests ($0.05/n$), where n represents the total number of SNPs used in the association analysis].

The TC and Hmp datasets were used for the heterosis analysis. The degree of dominance ' d/a ' was calculated using the peak SNPs of the heterosis-associated loci, where ' d ' and ' a ' represent the dominant and the additive effects, respectively. For each significant SNP detected in the TC and Hmp datasets, we denoted the allelic effect of the SNP estimated in the Hmp dataset as the dominant effect and that in the V dataset as the additive effect. This was similar to the method in our previous study (Li et al., 2008). The size of $|d/a|$ is the dominance degree, with $|d/a| > 1.2$ indicating overdominance (OD), $0.8 < |d/a| \leq 1.2$ indicating dominance (D), $0.2 < |d/a| \leq 0.8$ indicating partial dominance (PD) and $|d/a| \leq 0.2$ indicating additive (A). A GO (GeneOntology) analysis was conducted using candidate genes.

Superior alleles of the significant associated loci analysis.

The effects of heterozygous and homozygous genotypes were calculated for each peak SNP of the associated locus, and the average phenotypic measurements of the heterozygous and homozygous genotypes were calculated. For each significant associated locus, the genotype of the SNP with the preferred quality performance (such as lower AC and greater GL) was set as the superior gene allele (Huang et al., 2015). The numbers of superior alleles of each accession in each dataset were counted.

Sequence analysis

Because all the parental lines were re-sequenced using the *japonica* cultivar Nipponbare as the references, sequences of some known genes associated with quality traits, such as *OsACS6*, *Wx*, *ALK*, *GW5* and *GS3*, in male parental lines or testcross hybrids were obtained and aligned using the ClustalW method (Thompson et al., 2002).

Predicting ACs and yields of untested cross hybrids using BayesB method of genomic selection

Following Yang et al.'s method (2019), using the BayesB method, a twenty times fivefold cross-validation was conducted to evaluate the predictability (the correlation coefficient between the observed and predicted phenotypic values) of the rice hybrid AC and yield performances of 565 testcross hybrids using all the SNPs. According to the parameters estimated from the training sample (565 testcross hybrids), AC and yields of all the potential hybrids among the 113 male parental lines were predicted.

Authorship

LL, XZ, XZ performed data analyses and drafted the manuscript. KX, SS, JS, ZL, WT and ZT assisted in data analysis and discussion. YD, and XZ developed the populations. KX and CW assisted in field data collection. LL and ZY designed and finalized the manuscript, and revised and submitted manuscript. ZH conceived the project, planned, secured extramural funds.

Acknowledgements

This work was supported by funding from the National key technology research and development program (2016YFD0100101), Open Research Fund of State Key Laboratory of

441 Hybrid Rice (Hunan Hybrid Rice Research Center (2019KF05); Wuhan University
442 (KF201912)) , Research Foundation of Education Bureau of Hunan Province (19A244),
443 Double first-class construction project of Hunan Agricultural University (SYL2019028),
444 Hunan Science and Technology Plan Program (2019RS1055), Hubei Provincial cooperative
445 Innovation Center (Hubei Science and Education letter (2016) No.3.

446 We thank Prof. Sibin Yu and Tongmin Mou from Huazhong Agricultural University, Prof.
447 Shuangcheng Li from Sichuan Agricultural University, Prof. Shuzhu Tang from Yangzhou
448 University, Prof. Guanghua He from Southwest University, Prof. Huaxiong Qi from Hubei
449 Academy of Agricultural Sciences for providing material to us.

450

451 **Conflicts of interests**

452 The authors declare that there is no conflict of interests.

453

454 **Supplementary Figures**

455 **Supplementary Figure 1.** Phenotype distributions about general combining ability (Gca) of
456 twelve quality traits in parent lines.

457 **Supplementary Figure 2.** Phenotype distributions about mid-parent heterosis (Hmp) of twelve
458 quality traits in testcross lines.

459 **Supplementary Figure 3.** Phenotypic correlations among different traits. Correlations were
460 measured in Pearson correlation coefficients. The left bottom panel was phenotype
461 correlations in the TC dataset, and the right top panel was phenotype correlations in the Hmp
462 dataset. Asterisks indicate significant correlations using a two tailed t-test (* $P<0.05$, ** $P<0.01$

463 and *** $P < 0.001$).

464 **Supplementary Figure 4.** Genetic architecture of 120 parent lines. (a) Neighbor-joining tree
465 of 115 male parents and 5 female parents used to construct the NCII population. (b) Genetic
466 structure of parental lines analyzed using the program admixture.

467 **Supplementary Figure 5.** Genetic architecture of testcross lines. (a) Neighbor-joining tree of
468 565 testcross lines. (b) Genetic structure of testcross lines analyzed using the program
469 admixture.

470 **Supplementary Figure 6.** Plots of first two component in 120 parental lines.

471 **Supplementary Figure 7.** Plots of first four component in 565 testcross lines. (a) Plot of first
472 and second components in testcross lines; (b) Plot of third and fourth component in testcross
473 lines; (c) Plot of first three components in testcross lines.

474 **Supplementary Figure 8.** Manhattan plots of GL. (a) Manhattan plots of GL_V using four
475 different algorithms; (b) Manhattan plots of GL_Gca using four different algorithms; (c)
476 Manhattan plots of GL_TC using four different algorithms; (d) Manhattan plots of GL_Hmp
477 using four different algorithms.

478 **Supplementary Figure 9.** QQ-plots of GL. (a) QQ-plots of GL_V using four different
479 algorithms; (b) QQ-plots of GL_Gca using four different algorithms; (c) QQ-plots of GL_TC
480 using four different algorithms; (d) QQ-plots of GL_Hmp using four different algorithms.

481 **Supplementary Figure 10.** Manhattan plots of GW. (a) Manhattan plots of GW_V using four
482 different algorithms; (b) Manhattan plots of GW_Gca using four different algorithms; (c)
483 Manhattan plots of GW_TC using four different algorithms; (d) Manhattan plots of GW_Hmp
484 using four different algorithms.

Supplementary Figure 11. QQ-plots of GW. (a) QQ-plots of GW_V using four different algorithms; (b) QQ-plots of GW_Gca using four different algorithms; (c) QQ-plots of GW_TC using four different algorithms; (d) QQ-plots of GW_Hmp using four different algorithms.

Supplementary Figure 12. Manhattan plots of GLWR. (a) Manhattan plots of GLWR_V using four different algorithms; (b) Manhattan plots of GLWR_Gca using four different algorithms; (c) Manhattan plots of GLWR_TC using four different algorithms; (d) Manhattan plots of GLWR_Hmp using four different algorithms.

Supplementary Figure 13. QQ-plots of GLWR. (a) QQ-plots of GLWR_V using four different algorithms; (b) QQ-plots of GLWR_Gca using four different algorithms; (c) QQ-plots of GLWR_TC using four different algorithms; (d) QQ-plots of GLWR_Hmp using four different algorithms.

Supplementary Figure 14. Manhattan plots of CD. (a) Manhattan plots of CD_V using four different algorithms; (b) Manhattan plots of CD_Gca using four different algorithms; (c) Manhattan plots of CD_TC using four different algorithms; (d) Manhattan plots of CD_Hmp using four different algorithms.

Supplementary Figure 15. QQ-plots of CD. (a) QQ-plots of CD_V using four different algorithms; (b) QQ-plots of CD_Gca using four different algorithms; (c) QQ-plots of CD_TC using four different algorithms; (d) QQ-plots of CD_Hmp using four different algorithms.

Supplementary Figure 16. Manhattan plots of PGWC. (a) Manhattan plots of PGWC_V using four different algorithms; (b) Manhattan plots of PGWC_Gca using four different algorithms; (c) Manhattan plots of PGWC_TC using four different algorithms; (d) Manhattan plots of PGWC_Hmp using four different algorithms.

Supplementary Figure 17. QQ-plots of PGWC. (a) QQ-plots of PGWC_V using four different algorithms; (b) QQ-plots of PGWC_Gca using four different algorithms; (c) QQ-plots of PGWC_TC using four different algorithms; (d) QQ-plots of PGWC_Hmp using four different algorithms.

Supplementary Figure 18. Manhattan plots of TD. (a) Manhattan plots of TD_V using four different algorithms; (b) Manhattan plots of TD_Gca using four different algorithms; (c) Manhattan plots of TD_TC using four different algorithms; (d) Manhattan plots of TD_Hmp using four different algorithms.

Supplementary Figure 19. QQ-plots of TD. (a) QQ-plots of TD_V using four different algorithms; (b) QQ-plots of TD_Gca using four different algorithms; (c) QQ-plots of TD_TC using four different algorithms; (d) QQ-plots of TD_Hmp using four different algorithms.

Supplementary Figure 20. Manhattan plots of AC. (a) Manhattan plots of AC_V using four different algorithms; (b) Manhattan plots of AC_Gca using four different algorithms; (c) Manhattan plots of AC_TC using four different algorithms; (d) Manhattan plots of AC_Hmp using four different algorithms.

Supplementary Figure 21. QQ-plots of AC. (a) QQ-plots of AC_V using four different algorithms; (b) QQ-plots of AC_Gca using four different algorithms; (c) QQ-plots of AC_TC using four different algorithms; (d) QQ-plots of AC_Hmp using four different algorithms.

Supplementary Figure 22. Manhattan plots of ASV. (a) Manhattan plots of ASV_V using four different algorithms; (b) Manhattan plots of ASV_Gca using four different algorithms; (c) Manhattan plots of ASV_TC using four different algorithms; (d) Manhattan plots of ASV_Hmp using four different algorithms.

Supplementary Figure 23. QQ-plots of ASV. (a) QQ-plots of ASV_V using four different algorithms; (b) QQ-plots of ASV_Gca using four different algorithms; (c) QQ-plots of ASV_TC using four different algorithms; (d) QQ-plots of ASV_Hmp using four different algorithms.

Supplementary Figure 24. Manhattan plots of PC. (a) Manhattan plots of PC_V using four different algorithms; (b) Manhattan plots of PC_Gca using four different algorithms; (c) Manhattan plots of PC_TC using four different algorithms; (d) Manhattan plots of PC_Hmp using four different algorithms.

Supplementary Figure 25. QQ-plots of PC. (a) QQ-plots of PC_V using four different algorithms; (b) QQ-plots of PC_Gca using four different algorithms; (c) QQ-plots of PC_TC using four different algorithms; (d) QQ-plots of PC_Hmp using four different algorithms.

Supplementary Figure 26. Manhattan plots of BRR. (a) Manhattan plots of BRR_V using four different algorithms; (b) Manhattan plots of BRR_Gca using four different algorithms; (c) Manhattan plots of BRR_TC using four different algorithms; (d) Manhattan plots of BRR_Hmp using four different algorithms.

Supplementary Figure 27. QQ-plots of BRR. (a) QQ-plots of BRR_V using four different algorithms; (b) QQ-plots of BRR_Gca using four different algorithms; (c) QQ-plots of BRR_TC using four different algorithms; (d) QQ-plots of BRR_Hmp using four different algorithms.

Supplementary Figure 28. Manhattan plots of MRR. (a) Manhattan plots of MRR_V using four different algorithms; (b) Manhattan plots of MRR_Gca using four different algorithms; (c) Manhattan plots of MRR_TC using four different algorithms; (d) Manhattan plots of

551 MRR_Hmp using four different algorithms.

552 **Supplementary Figure 29.** QQ-plots of MRR. (a) QQ-plots of MRR_V using four different

553 algorithms;

554 (b) QQ- plots of MRR_Gca using four different algorithms; (c) QQ-plots of MRR_TC using

555 four different algorithms; (d) QQ-plots of MRR_Hmp using four different algorithms.

556 **Supplementary Figure 30.** Manhattan plots of HRR. (a) Manhattan plots of HRR_V using

557 four different algorithms; (b) Manhattan plots of HRR_Gca using four different algorithms; (c)

558 Manhattan plots of HRR_TC using four different algorithms; (d) Manhattan plots of HRR_Hmp

559 using four different algorithms.

560 **Supplementary Figure 31.** QQ-plots of HRR. (a) QQ-plots of HRR_V using four different

561 algorithms; (b) QQ- plots of HRR_Gca using four different algorithms; (c) QQ-plots of

562 HRR_TC using four different algorithms; (d) QQ-plots of HRR_Hmp using four different

563 algorithms.

564 **Supplementary Figure 32.** The dominance degree of 12 quality traits. (a) The dominance of

565 V in each trait; (b) The dominance of Gca in each trait; (c) The dominance of TC in each trait;

566 (d) The dominance of Hmp in each trait.

567 **Supplementary Figure 33.** LD decay in 120 parental lines.

568 **Supplementary Figure 34.** Analyses of the peak for amylose content (near gene *Wx*)

569 associated haplotypes at chromosome 6 in TC. Local Manhattan plot (top) and LD heatmap

570 (bottom) surrounding the peak on chromosome 6. The red dash line indicated the positions of

571 of *Wx*.

572 **Supplementary Figure 35.** Analyses of the peak for amylose content (near gene *OsACS6s*)

associated haplotypes at chromosome 6 in V (a) and TC dataset (b). Local Manhattan plot (top) and LD heatmap (bottom) surrounding the peak on chromosome 6. The red dash line indicated the positions of of *OsACS6s*.

Supplementary Figure 36. Analyses of the peak for amylose content (near gene *ALK*) associated haplotypes at chromosome 6 in V (a) and TC dataset (b). Local Manhattan plot (top) and LD heatmap (bottom) surrounding the peak on chromosome 6. The red dash line indicated the positions of of *ALK*.

Supplementary Figure 37. A high degree of polymorphism in the *OsACS6* coding sequence, and its haplotypes in V and TC datasets. (a) Haplotype analysis of *OsACS6*. (b) Boxplots for grain width based on the haplotypes of *OsACS6*.

Supplementary Figure 38. A high degree of polymorphism in the *ALK* coding sequence, and its haplotypes in V and TC datasets. (a) Haplotype analysis of *ALK*. (b) Boxplots for grain width based on the haplotypes of *ALK*.

Supplementary Figure 39. A high degree of polymorphism in the *GW5* coding sequence, and its haplotypes in V and TC datasets. (a) Haplotype analysis of *GW5*. (b) Boxplots for grain width based on the haplotypes of *GW5*.

Supplementary Figure 40. A high degree of polymorphism in the *GS3* coding sequence, and its haplotypes in V and TC datasets. (a) Haplotype analysis of *GS3*. (b) Boxplots for grain width based on the haplotypes of *GS3*. (c) Unrooted phylogenetic tree of *GS3*.

Supplementary Tables

595 **Supplementary Table 1.** The information of 115 varieties and 5 sterile lines used in NCII
596 design.

597 **Supplementary Table 2.** Selected model of GWAS for 12 rice quality traits.

598 **Supplementary Table 3.** The information of SNP pleiotropism.

599 **Supplementary Table 4.** The average value of each trait in different number of Superior alleles.

600 **Supplementary Table 5.** Untested cross AC performance and their SNP genotype of
601 associated loci or haplotype of gene.

602 **Supplementary Table 6.** Untested cross GL performance and their SNP genotype of associated
603 loci or haplotype of gene.

604 **Supplementary Table 7.** Untested cross GW performance and their SNP genotype of
605 associated loci or haplotype of gene.

606 **Supplementary Table 8.** Phenotype data of parental lines.

607 **Supplementary Table 9.** Phenotype data of testcross lines.

608

609 **References**

610 Alexander, D., Novembre J., Lange K. (2009) Fast model-based estimation of ancestry in
611 unrelated individuals. *Genome Res*, 19(9):1655-1664.

612 Bao, J. (2014). Genes and QTLs for rice grain quality improvement. *InTech-Open Science*
613 *Open Mind*, 239-278.

614 Che, R., Tong, H., Shi, B., Liu, Y., Fang, S., Liu, D., ... & Zhao, M. (2016). Control of grain
615 size and rice yield by GL2-mediated brassinosteroid responses. *Nature plants*, 2(1), 15195.

616 Chen, J., Zhou, H., Xie, W., Xia, D., Gao, G., Zhang, Q., ... & He, Y. (2019). Genome-wide
617 association analyses reveal the genetic basis of combining ability in rice. *Plant biotechnology*
618 *journal*. <https://doi.org/10.1111/pbi.13134>

619 Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., ... & Pan, Y. (2017). Genome-wide
620 association studies dissect the genetic networks underlying agronomical traits in soybean.
621 *Genome biology*, 18(1), 161. DOI 10.1186/s13059-017-1289-9

622 Fitzgerald, M. A., & Resurreccion, A. P. (2009). Maintaining the yield of edible rice in a
623 warming world. *Functional Plant Biology*, 36(12), 1037-1045.

624 Han, X., Wang, Y., Liu, X., Jiang, L., Ren, Y., Liu, F., ... & Wang, J. (2011). The failure to
625 express a protein disulphide isomerase-like protein results in a floury endosperm and an
626 endoplasmic reticulum stress response in rice. *Journal of experimental botany*, 63(1), 121-130.

627 Huang, X., & Han, B. (2018). Genome-wide association mapping of complex traits in rice. In
628 *Rice genomics, genetics and breeding* (pp. 497-510). Springer, Singapore.

629 Huang, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., ... & Fan, D. (2010). Genome-wide
630 association studies of 14 agronomic traits in rice landraces. *Nature genetics*, 42(11), 961.

631 Han, B., (2010). Genome-wide association studies of 14 agronomic traits in rice
632 landraces. *Nature genetics*, 42(11), 961.

633 Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., ... & Chen, N. (2016). Genomic
634 architecture of heterosis for yield traits in rice. *Nature*, 537(7622), 629.

635 Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., ... & Chen, N. (2015). Genomic

636 analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis.

637 Nature communications, 6, 6258.

638 Huang, X., Zhao, Y., Li, C., Wang, A., Zhao, Q., Li, W., ... & Lu, Y. (2012). Genome-wide
639 association study of flowering time and grain yield traits in a worldwide collection of rice
640 germplasm. *Nature genetics*, 44(1), 32-39.

641 Huggins, T. D., Chen, M. H., Fjellstrom, R. G., Jackson, A. K., McClung, A. M., & Edwards,
642 J. D. (2018). Association analysis of three diverse rice (*Oryza sativa* L.) germplasm collections
643 for loci regulating grain quality traits. *The Plant Genome*.

644 Ishimaru, K., Hirotsu, N., Madoka, Y., Murakami, N., Hara, N., Onodera, H., ... & Miyagawa,
645 H. (2013). Loss of function of the IAA-glucose hydrolase gene *TGW6* enhances rice grain
646 weight and increases yield. *Nature genetics*, 45(6), 707.

647 Kang, H. G., Park, S., Matsuoka, M., & An, G. (2005). White-core endosperm floury
648 endosperm-4 in rice is generated by knockout mutations in the C4-type pyruvate
649 orthophosphate dikinase gene (*OsPPDKB*). *The Plant Journal*, 42(6), 901-911.

650 Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C.,
651 Eskin, E. (2010). Variance component model to account for sample structure in genome-wide
652 association studies. *Nat Genet*, 42(4):348-354.

653 Lee, T. H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: a pipeline to
654 construct a phylogenetic tree from huge SNP data. *BMC genomics*, 15(1), 162.

655 Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler

656 Transform. Bioinformatics, 25:1754-60.

657 Li, L. , Lu, K. , Chen, Z. , Mu, T. , Hu, Z. , & Li, X. . (2008). Dominance, overdominance and
658 epistasis condition the heterosis in two heterotic rice hybrids. *Genetics*, 180(3), 1725-1742.

659 Li, T., Ma, X., Li, N., Zhou, L., Liu, Z., Han, H., ... & Dai, X. (2017). Genome-wide association
660 study discovered candidate genes of Verticillium wilt resistance in upland cotton (*Gossypium*
661 *hirsutum* L.). *Plant biotechnology journal*, 15(12), 1520-1532.

662 Li, Y., Fan, C., Xing, Y., Jiang, Y., Luo, L., Sun, L., ... & He, Y. (2011). Natural variation in
663 GS5 plays an important role in regulating grain size and yield in rice. *Nature genetics*, 43(12),
664 1266.

665 Li, Y., Fan, C., Xing, Y., Yun, P., Luo, L., Yan, B., ... & Xiao, J. (2014). Chalk5 encodes a
666 vacuolar H⁺-translocating pyrophosphatase influencing grain chalkiness in rice. *Nature*
667 *genetics*, 46(4), 398.

668 Liao, P. F., Ouyang, J. X., Zhang, J. J., Yang, L., Wang, X., Peng, X. J., ... & Li, S. B. (2019).
669 OsDCL3b affects grain yield and quality in rice. *Plant molecular biology*, 99(3), 193-204.

670 Liu, X., Huang, M., Fan, B., Buckler, E. S., Zhang, Z.. (2016). Iterative usage of fixed and
671 random effect models for powerful and efficient genome-wide association studies. *Plos*
672 *Genetics*, 12(2), e1005767.

673 Lou, J. , Chen, L. , Yue, G. , Lou, Q. , Mei, H. , & Xiong, L. , et al. (2009). QTL mapping of
674 grain quality traits in rice. *Journal of Cereal Science*, 50(2), 145-151.

675 Mao, H., Sun, S., Yao, J., Wang, C., Yu, S., Xu, C., ... & Zhang, Q. (2010). Linking differential
676 domain functions of the GS3 protein to natural variation of grain size in rice. *Proceedings of*
677 *the National Academy of Sciences*, 107(45), 19579-19584.

678 Misra, G., Badoni, S., Domingo, C. J., Cuevas, R. P. O., Llorente, C., Mbanjo, E. G. N.,
679 Sreenivasulu, N.. (2018). Deciphering the Genetic Architecture of Cooked Rice Texture. *Front*
680 *Plant Science*. 9:1405. doi: 10.3389/fpls.2018.01405.

681 McCouch, S. R., Zhao, K., Wright, M., Tung, C. W., Ebana, K., Thomson, M., Reynolds, A.
682 Wang, D., Declerck, G., Ali, M. L., McClung, A., Eizenga, G., Bustamante, C..
683 (2010). Development of genome-wide SNP assays for rice. *Breeding Science*, 60(5), 524-535.

684 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J.,
685 Sklar, P., de Bakker, P. I., Daly, M. J., Sham, P. C. (2007) PLINK: a tool set for whole-genome
686 association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559-575.

687 Qu, Z. , Li, L. , Luo, J. , Wang, P. , Yu, S. , & Mou, T. , et al. (2012). Correction: qtl mapping
688 of combining ability and heterosis of agronomic traits in rice backcross recombinant inbred
689 lines and hybrid crosses. PLOS ONE, 7.

690 Ren, D., Rao, Y., Huang, L., Leng, Y., Hu, J., Lu, M., ... & Guo, L. (2016). Fine mapping
691 identifies a new QTL for brown rice rate in rice (*Oryza Sativa* L.). *Rice*, 9(1), 4.

692 Roberts, A., McMillan, L., Wang, W., Parker, J., Rusyn, I., & Threadgill, D. (2007). Inferring
693 missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding
694 windows. *Bioinformatics*, 23(13), i401-i407.

695 Smith, A. et al. (1997) The synthesis of the starch granule. *Annu. Rev. Plant Physiol. Plant Mol.*
696 *Biol.*, 48, 67-87

697 Song, X. J., Huang, W., Shi, M., Zhu, M. Z., & Lin, H. X. (2007). A QTL for rice grain width
698 and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature*
699 *genetics*, 39(5), 623.

700 Song, X. J., Kuroha, T., Ayano, M., Furuta, T., Nagai, K., Komeda, N., ... & Suzuki, T. (2015).
701 Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight,
702 yield, and plant biomass in rice. *Proceedings of the National Academy of Sciences*, 112(1), 76-
703 81.

704 Sun, L., Li, X., Fu, Y., Zhu, Z., Tan, L., Liu, F., ... & Sun, C. (2013). GS 6, A Member of the
705 GRAS Gene Family, Negatively Regulates Grain Size in Rice. *Journal of integrative plant*
706 *biology*, 55(10), 938-949.

707 Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z. B., Pang Y. C., Liu, D., Lipkai,
708 A. E., Buckler, E. S., Zhang, Z. W. (2016). GAPIT version 2: an enhanced integrated tool for
709 genomic association and prediction. *The plant genome*, 9(2) doi:
710 10.3835/plantgenome2015.11.0120.

711 Thompson, J. D. , Gibson, T. J. , & Higgins, D. G. . (2003). Multiple sequence alignment using
712 clustalw and clustalx. *Current Protocols in Bioinformatics*, 00(1).

713 Tung, C. W., Zhao, K., Wright, M. H., Ali, M. L., Jung, J., Kimball, J., Tyagi, W., Thomson,
714 M. J., McNally, K., Leung, H., Kim, H., Ahn, S-N., Reynolds, A., Scheffler, B., Eizenga,

715 G., Mcclung, A., Bustamante, C., Mccouch, S. R.. (2010). Development of a research
716 platform for dissecting phenotype–genotype associations in rice (*Oryza* spp.). *Rice*, 3(4), 205-
717 217.

718 Wang, E., Wang, J., Zhu, X., Hao, W., Wang, L., Li, Q., ... & Ma, H. (2008). Control of rice
719 grain-filling and yield by a gene with a potential signature of domestication. *Nature*
720 *genetics*, 40(11), 1370.

721 Wang, S., Li, S., Liu, Q., Wu, K., Zhang, J., Wang, S., ... & Wang, F. (2015). The OsSPL16-
722 GW7 regulatory module determines grain shape and simultaneously improves rice yield and
723 grain quality. *Nature genetics*, 47(8), 949.

724 Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., Lin, X., ... & Zhang, G. (2012). Control of grain
725 size, shape and quality by OsSPL16 in rice. *Nature genetics*, 44(8), 950.

726 Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., Xu, J., ... & Ye, W. (2015). Copy number
727 variation at the GL7 locus contributes to grain size diversity in rice. *Nature genetics*, 47(8),
728 944.

729 Weng, J., Gu, S., Wan, X., Gao, H., Guo, T., Su, N., ... & Wang, J. (2008). Isolation and initial
730 characterization of GW5, a major QTL associated with rice grain width and weight. *Cell*
731 *research*, 18(12), 1199.

732 Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., ... & Zhang, Q. (2010). Parent-
733 independent genotyping for constructing an ultrahigh-density linkage map based on population
734 sequencing. *Proceedings of the National Academy of Sciences*, 107(23), 10578-10583.

735 Xu, F., Fang, J., Ou, S., Gao, S., Zhang, F., Du, L., ... & Wang, G. (2015). Variations in CYP
736 78 A 13 coding region influence grain size and yield in rice. *Plant, cell & environment*, 38(4),
737 800-811.

738 Zhang, X., Wang, J., Huang, J., Lan, H., Wang, C., Yin, C., ... & Zhang, H. (2012). Rare allele
739 of OsPPKL1 associated with grain length causes extra-large grain and a significant yield
740 increase in rice. *Proceedings of the National Academy of Sciences*, 109(52), 21534-21539.

741 Zhao, D. S., Li, Q. F., Zhang, C. Q., Zhang, C., Yang, Q. Q., Pan, L. X., ... & Liu, Q. Q. (2018).
742 GS9 acts as a transcriptional activator to regulate rice grain shape and appearance
743 quality. *Nature communications*, 9(1), 1240.

744 Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J.,
745 Islam, M. R., Reynolds, A., Mezey, J., McClung, A. M., Bustamante, C. D., McClung, A. M.
746 (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits
747 in *Oryza sativa*. *Nature communications*, 2, 467.

748 Zhen, G. , Qin, P. , Liu, K. Y. , Nie, D. Y. , Yang, Y. Z. , & Deng, X. W. , et al. (2017). Genome-
749 wide dissection of heterosis for yield traits in two-line hybrid rice populations. *Scientific*
750 *Reports*, 7(1), 7635.

751 Zhou, H., Wang, L., Liu, G., Meng, X., Jing, Y., Shu, X., ... & Wu, D. (2016). Critical roles of
752 soluble starch synthase SSIIIa and granule-bound starch synthase Waxy in synthesizing
753 resistant starch in rice. *Proceedings of the National Academy of Sciences*, 113(45), 12844-
754 12849.

Table 1 Significant markers associated with rice grain quality traits in four datasets

Dataset	Trait	SNP	Chromosome	Position	Effect	<i>P</i> value	Dominance Degree	GeneID	Gene Symbol
V	AC	chr06_1629652	6	1629652	-3.46	2.55E-24	A	Os06g0130400	OsACS6; SSG6
V	AC	chr06_23373922	6	23373922	-3.95	2.16E-17	A	Os06g0594400	
V	AC	chr01_8798760	1	8798760	2.58	1.55E-12	A	Os01g0261200	
V	AC	chr01_13792415	1	13792415	1.51	9.30E-10	A	Os01g0347000	
V	GL	chr01_19919075	1	19919075	-0.22	1.07E-11	A	Os01g0541600	
V	GL	chr03_16761936	3	16761936	0.17	2.23E-09	A	Os03g0407900	

V	GL	chr05_20726376	5	20726376	0.17	1.42E-08	A	Os05g0422300	qSW5; GW5
V	GLWR	chr02_19522464	2	19522464	0.16	2.70E-18	A	Os02g0530600	
V	GLWR	chr05_5359598	5	5359598	0.16	1.66E-14	A	Os05g0187500	
V	GLWR	chr05_18420108	5	18420108	-0.1	5.53E-12	A	Os05g0381400	
V	GLWR	chr08_19397555	8	19397555	0.21	3.41E-10	PD	Os08g0405700	
V	GLWR	chr11_18104419	11	18104419	-0.08	2.36E-08	A	Os11g0507300	
V	GW	chr10_17726899	10	17726899	-0.11	4.64E-14	A	Os10g0476000	
Gca	AC	chr06_1621853	6	1621853	-1.64	5.57E-25	PD	Os06g0130200	
Gca	AC	chr02_9560639	2	9560639	1.67	6.74E-13	PD	Os02g0267200	
Gca	AC	chr06_20096769	6	20096769	-1.56	2.13E-12	OD	Os06g0536333	
Gca	AC	chr01_869499	1	869499	0.94	3.18E-12	A	Os01g0116100	

Gca	AC	chr08_20841182	8	20841182	0.7	4.56E-09	PD	Os08g0431100
Gca	CD	chr02_7089314	2	7089314	4.43	1.33E-12	A	Os02g0226500
Gca	CD	chr12_7386283	12	7386283	-1.96	4.23E-10	A	Os12g0234800
Gca	CD	chr02_34537381	2	34537381	-2.1	1.08E-09	A	Os02g0808100
Gca	GL	chr03_16720836	3	16720836	-0.17	1.33E-19	OD	Os03g0407100
Gca	GL	chr02_22854719	2	22854719	0.08	1.79E-13	PD	Os02g0591500
Gca	GLWR	chr03_16720836	3	16720836	-0.14	1.97E-22	PD	Os03g0407100
Gca	GLWR	chr11_4766580	11	4766580	-0.06	3.51E-14	D	Os11g0195100
Gca	GLWR	chr09_14812304	9	14812304	-0.05	6.94E-12	PD	Os09g0414600
Gca	GLWR	chr05_5359838	5	5359838	-0.06	5.77E-11	A	Os05g0187500
Gca	GLWR	chr05_19787200	5	19787200	0.07	2.78E-10	PD	Os05g0406150

qSW5;
GW5

Gca	GW	chr05_5359838	5	5359838	0.05	1.73E-17	A	Os05g0187500	qSW5; GW5
Gca	GW	chr11_27108114	11	27108114	0.04	5.57E-16	PD	Os11g0670800	
Gca	GW	chr10_22867856	10	22867856	0.024913	2.01E-08	A	Os10g0575200	
Gca	PGWC	chr05_4140727	5	4140727	5.559854	2.67E-13	A	Os05g0169100	
Gca	PGWC	chr02_34531187	2	34531187	-4.38526	2.24E-11	A	Os02g0808100	
Gca	PGWC	chr02_35252088	2	35252088	5.31925	3.91E-10	A	Os02g0820900	
Gca	PGWC	chr12_21824818	12	21824818	-4.18341	1.58E-08	PD	Os12g0543101	
TC	AC	chr06_1731808	6	1731808	-4.42699	1.01E-22	A	Os06g0132300	OsACS6; SSG6
TC	CD	chr03_35239012	3	35239012	-6.98646	1.50E-08	PD	Os03g0838400	
TC	GL	chr03_35239012	3	35239012	-6.98646	1.50E-08	A	Os03g0838400	

TC	GL	chr04_31432541	4	31432541	0.22011	5.96E-18	PD	Os04g0618950	qSW5; GW5
TC	GLWR	chr03_16721158	3	16721158	0.111735	1.92E-14	PD	Os03g0407100	
TC	GW	chr05_5359838	5	5359838	0.039716	7.82E-11	A	Os05g0187500	
TC	GW	chr04_1164552	4	1164552	0.0406	1.34E-09	PD	Os04g0119800	
Hmp	AC	chr07_26451926	7	26451926	3.762949	1.39E-09	OD	Os07g0636800	OsPDK1
Hmp	ASV	chr02_23164894	2	23164894	-0.79736	1.07E-09	OD	Os02g0596500	
Hmp	ASV	chr05_10666	5	10666	-0.12458	2.32E-08	PD	Os05g0100008	
Hmp	BRR	chr07_26080972	7	26080972	-3.20805	1.78E-08	OD	Os07g0629000	
Hmp	CD	chr07_26490098	7	26490098	7.211639	2.58E-11	D	Os07g0637300	
Hmp	CD	chr02_27849011	2	27849011	-5.18957	1.23E-09	PD	Os02g0681900	
Hmp	CD	chr10_14574830	10	14574830	-4.23373	1.10E-08	OD	Os10g0416200	

Hmp	GL	chr03_7504625	3	7504625	0.131167	1.84E-09	OD	Os03g0241900
Hmp	GL	chr03_16678042	3	16678042	-0.09303	1.48E-08	OD	Os03g0406900
Hmp	GLWR	chr08_19397555	8	19397555	-0.14824	3.88E-10	PD	Os08g0405700
Hmp	GLWR	chr04_31298826	4	31298826	-0.02163	3.12E-09	OD	Os04g0616500
Hmp	GW	chr04_13899992	4	13899992	0.045955	1.87E-08	OD	Os04g0307800
Hmp	PGWC	chr07_26436340	7	26436340	6.413725	2.09E-11	PD	Os07g0636400
Hmp	PGWC	chr12_3127802	12	3127802	9.871879	1.72E-10	OD	Os12g0162100
Hmp	TD	chr09_22390062	9	22390062	0.627327	1.38E-11	OD	Os09g0563300
Hmp	TD	chr02_35240070	2	35240070	0.274355	2.42E-10	OD	Os02g0820700
Hmp	TD	chr03_27427059	3	27427059	-0.38817	3.85E-10	PD	Os03g0687200

Figure legends

Figure 1. Phenotype distributions and phenotypic correlations among different traits (a)

Phenotype distributions of rice grain quality traits in Varieties (V) and TC (Testcross) dataset.

(b) Phenotypic correlations among different traits. The left bottom panel represents the phenotypic correlations in the V dataset, and the right top panel represents the phenotypic correlations in the Gca dataset. The values indicate correlation coefficients (r). The areas and colors of ellipses represent the absolute values of the corresponding r . Right and left oblique ellipses indicated positive and negative correlations, respectively. Asterisks indicate significant correlations using a two tailed t -test (* $p < 0.05$, ** $p < 0.01$ and *** $p < 0.001$).

Figure 2. GWAS results for GL. (a, b) Manhattan plots and and quantile–quantile about V of GL. (c, d) Manhattan plots and and quantile–quantile about Gca of GL. (e, f) Manhattan plots and and quantile–quantile about TC of GL. (a, b) Manhattan plots and and quantile–quantile about Hmp of GL. The genome-wide significant P value threshold $P < 2.64 \times 10^{-8}$ is indicated by a horizontal dash-dot line. Significant SNPs with $P < 2.64 \times 10^{-8}$ are depicted as red dots. The well-characterized genes near significant loci are indicated in red characters.

Figure 3. GWAS results for AC. (a, b) Manhattan plots and and quantile–quantile about V of AC. (c, d) Manhattan plots and and quantile–quantile about Gca of AC. (e, f) Manhattan plots and and quantile–quantile about TC of AC. (a, b) Manhattan plots and and quantile–quantile about Hmp of AC. The genome-wide significant P value threshold $P < 2.64 \times 10^{-8}$ is indicated by a horizontal dash-dot line. Significant SNPs with $P < 2.64 \times 10^{-8}$ are depicted as red dots. The

well-characterized genes near significant loci are indicated in red characters.

Figure 4. The phenotype values and the percentage of accessions with different number of superior alleles of significant loci or favorable haplotype of closest gene of significant loci for quality trait in V and TC datasets. a, c, e separately represents AC, GW and GL with different number of superior alleles or favorable haplotype of genes in V dataset, b, d, f separately represents AC, GW and GL with different number of superior alleles or favorable haplotype of genes in TC dataset. Percentage showed on bars are rate of corresponding lines to whole lines in V or TC dataset.

Figure 5 A high degree of polymorphism in the *Wx* coding sequence, and its haplotypes in V and TC datasets. (a) Haplotype analysis of *Wx*. (b) Boxplots for grain width based on the haplotypes of *Wx*.

bioRxiv preprint doi: <https://doi.org/10.1101/2020.06.04.134023>; this version posted June 4, 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 4.0 International license.

Figure 2

Figure 3

a

b

Figure 1

a *Wx* Coding region
GT1 Glycogen synthase DULL1

1830 bp

Trait	Dataset	Hap	Pos	666	1078	1234	No.	mean±sd	YD.mean±sd
AC	V	Hap.A		A	T	C	33	18.66±6.21	32.09±11.08
		Hap.B		A	C	C	22	22.09±4.55	31.62±12.06
		Hap.C		A	C	T	49	18.50±5.50	26.15±10.09
		Hap.D		A	C	Y	2	13.16±0.09	32.86±9.88
		Hap.E		A	C	--	1	2.967	15.78
		Hap.F		C	C	C	1	26.051	34.05
		Hap.G		A	Y	C	2	19.83±5.96	30.13±1.03
		Hap.H		A	Y	Y	2	22.66±4.62	28.9±20.15
		Hap.I		C	T	C	1	14.054	26.73
	TC	Hap.J		A	T	C	165	19.56±4.49	43.24±13.14
		Hap.K		A	Y	C	10	19.58±4.45	40.12±12.57
		Hap.L		A	Y	Y	245	18.94±4.38	41.25±14.94
		Hap.M		A	Y	C/Y	10	15.92±3.77	40.50±6.86
		Hap.N		A	Y	-C	5	12.58±5.59	46.02±6.68
		Hap.O		M	Y	C	5	22.87±2.10	44.17±17.21
		Hap.P		A	Y/T	C	10	19.80±3.93	39.15±25.21
		Hap.Q		A	Y/T	C/Y	10	20.03±3.78	33.87±10.79
		Hap.R		M	T	C	5	17.65±4.35	39.65±5.86

Figure 5

Figure 4