Identification of candidate genes for gelatinization temperature, gel consistency and pericarp color by GWAS in rice based on SLAF-sequencing

Xinghai Yang[1], Xiuzhong Xia[1], Yu Zeng[1], Baoxuan Nong[1], Zongqiong Zhang[1], Yanyan Wu[2], Faqian Xiong[3], Yuexiong Zhang[1], Haifu Liang[1], Guofu Deng[1], Danting Li[1]

1 Rice Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China

2 Biotechnology Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China

3 Cash Crops Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China

Xinghai Yang and Xiuzhong Xia are co-first author.

Corresponding author: Danting Li

Address: 174 East Daxue Road, Nanning, Guangxi, 530007, China

Tel.:+867713243019

E-mail: ricegl@163.com.

Abstract

Rice is an important cereal in the world, uncovering the genetic basis of agronomic traits in rice landraces genes associated with agronomically important traits is indispensable for both understanding the genetic basis of phenotypic variation and efficient crop improvement. Gelatinization temperature, gel consistency and pericarp color are important indices of rice cooking and eating quality evaluation and potential nutritional importance, which attract wide attentions in the application of genetic and breeding. To dissect the genetic basis of gelatinization temperature (GT), gel consistency (GC) and pericarp color (PC), a total of 419 rice landraces core germplasm collections consisting of 330 *indica* lines, 78 *japonica* lines and 11 uncertain varieties were grown, collected, then GT, GC, PC were measured for two years, and sequenced using Specific Locus Amplified Fragment Sequencing (SLAF) technology. In this study, 261,385,070 clean reads and 56,768 polymorphic SLAF tags were obtained, which a total of 211,818 single nucleotide polymorphisms (SNPs) were discovered. With 208,993 SNPs meeting the criterion of minor allele frequency (MAF) > 0.05 and integrity> 0.5, the phylogenetic tree and population structure analysis were performed for all 419 rice landraces, and the whole panel mainly separated into six subpopulations based on population structure analysis. Genome-wide association study (GWAS) was carried out for the whole panel, *indica* subpanel and *japonica* subpanel with subset SNPs respectively. One quantitative trait locus (QTL) on chromosome 6 for GT was detected in the whole panel and *indica* subpanel, and one QTL associated with GC was located on chromosome 6 in the whole panel and *indica* subpanel. For the PC trait, 8 QTLs were detected in the whole panel on chromosome 1, 3, 4, 7, 8, 10 and 11, and 7 QTLs in the *indica* subpanel on chromosome 3, 4, 7, 8, 10 and 11. The loci on chromosome 3, 8, 10 and 11 have not been identified previously, and they may be the candidate genes of pericarp color. For the three traits, no QTL was detected in *japonica* subpanel，probably because of the polymorphism repartition between the subpanel, or small population size of *japonica* subpanel. This paper provides new gene resources and insights into the molecular mechanisms of important agricultural trait of rice phenotypic variation and genetic improvement of rice quality variety breeding.

Introduction

Rice is one of the most important food crops in the world[1], the increased rice yield, improved quality and advanced resistance to biotic and abiotic stress play an important role in solving the world food problem, improving people's life quality and reducing environmental pollution. Identification and utilization of favorable genes in rice germplasm resources is the foundation of rice breeding. Guangxi is likely the origin of cultivated rice[2], which possesses a large number of rice landraces containing rich natural variation and genetic diversity, and it is the important genetic resources for breed improvement.

Quantitative trait loci (QTLs) mapping has been widely used to explore the

genetic basis of complex agronomic traits in different crops. So far, yield-related genes [3-6], quality-related genes [7-9], and resistance-related genes [10-12] and so on have been identified and cloned through biparental linkage mapping in rice (http://www.gramene.org/, http://www.ricedata.cn/index.htm). However, almost studies need to construct mapping populations (e.g. F_2, BC, RILs, NILs), which is very time-consuming and painstaking [13].

In recent years, with the rapid development of high-throughput sequencing technology and reduced sequencing cost, genome-wide association study (GWAS) based on SNPs has become a new method for studying important agronomic traits in rice [14-20], maize [21], sorghum [22], soybean [23], tomato [24] , et al. Previous studies have documented that GWAS could be a useful tool to dissect the genetic changes for complicated traits has some advantages. First, GWAS can efficiently detect multiple QTLs in the same population meanwhile; Second, the associated population includes the majority of variations of the related loci. Therefore, the GWAS not only can help us understand the gene function, but also discover the favorable alleles for genetic improvement of plants. However, GWAS has some limitations. (i) the large effect variations and minor effect genes can not be identified easily by GWAS [25]; (ii) genetic heterogeneity can reduce the efficacy of mutation detection[25,26]; (iii) the population structure lead to false positive associations between phenotype and unlinked markers [18]; (iv) GWAS can be limited by the genetic characteristics of different species, for example, the LD decay of rice is lower than that of outcrossing maize [27], so the GWAS still can not replace the traditional map based clone to achieve fine mapping of the target gene [18,28].

Although a number of GWAS analysis for important agronomic traits in rice have been performed, new QTLs and candidate genes can be found through different populations [14-20]. In this study, we used 419 rice landraces core germplasm collections from Guangxi to explore the molecular basis of GT, GC and PC, the panel was genotyped based on SLAF-seq technology, the three traits were measured for twice during 2014 and 2015, and then the GWAS was performed. The research provided genetic resources for molecular breeding and shed light on rice quality improvement.

Materials and methods

Plant materials and phenotyping

The diversity panel was composed of 419 landraces collected from core rice germplasm of Guangxi. The lines were planted in Nanning experimental field (China at 22.85°N, 108.26°E) from July 2014 to November 2014 and from July 2015 to November 2015. For each landrace, five randomly chosen plants were harvested when they were mature and used for measurement of gelatinization temperature and gel consistency and recording the pericarp color. The gelatinization temperature was measured by the alkali digestion test [29] and the gel consistency was estimated based on Cagampang et al. (1973) [30].

SLAF sequencing and SNP genotyping

Total genomic DNA was extracted from young quadrifoliate leaves of all rice landraces using cetyltrimethyl annonium bromide (CTAB) protocol [31], and digested by two restriction enzymes RsaI and HaeIII. The SLAF sequencing was performed on an IlluminaHiseq 2500 system. The polymorphic SLAF tags were obtained by clustering the clean reads using BLAT software [32] , aligned to reference genome (*Oryza sativa* L. spp . *japonica*. cv. Nipponbare, http://plants.ensembl.org/index.html) using the BWA software [33], and then the SNP calling was performed using GATK [34] and Samtools packages [35]. A total of 208,993 SNPs with a minor allele frequency (MAF)>0.05 and integrity > 0.5 was retained for GWAS.

Genetic kinship calculation, phylogenetic tree construction, principal component analysis, population sturcture

Based on the 208,993 high quality SNPs, the phylogenetic tree was constructed by MEGA5 [36]. the pairwise kinship was carried out using the SPAGeDi software package [37], and the population structure was analysed by Admixture software [38], which the subpanel number was predicted from 1 to 10. The kinship and population structure analysis was performed for the whole panel, *indica* subpanel and *japonica* subpanel respectively in convenience of subsequent GWAS for the three combinations. Principal component analysis (PCA) was performed using GAPIT [39].

Genome-wide association study

In this study, in order to eliminate the effort of population structure between *indica* subpopulation and *japonica* subpopulation, GWAS was proceeded for the whole panel, *indica* subpanel and *japonica* subpanel respectively using the mixed linear model (MLM) of Tassel v3 [40], which took the population structure and kinship into consideration, and the significant P value was set to 4.79×10^{-8}.

Result

Phenotypic variation of gelatinization temperature, gel consistency and pericarp color

The alkali spreading value can be used to measure the gelatinization temperature , which is inversely related to GT, and ranged from 0 to 6.15 with an average of 2.86 in 2014, and 0 to 7 with an average of 3.74 in 2015; The gel consistency spanned 26 to 100 with an average of 67.91 in 2014, and 6 to 100 with an average of 70.94 in 2015; for the pericarp color trait, the number of rice landraces with white, red, black color is 308, 97, 14 respectively (S1 Table).

Analysis of SLAF-seq data and development of SNPs

After sequencing data quality control, a total of 67,665 SLAF tags were obtained with an average sequencing depth of 8.75×. Ultimately, 56,768 polymorphic SLAF tags were retained when aligned to the reference genome (S2 Table).

A total of 211,818 SNPs was identified using the GATK and samtools software package, of these, the number of SNPs with MAF > 0.05 and integrity> 0.5 was

208,993, which distributed on every chromosome with a mean number 17,416 per chromosome. For the high quality SNPs, 104,068 SNPs located in gene region, and 107750 SNPs in intergenic region. The chromosome 10 possessed the most SNP density while chromosome 2 for the fewest SNP density (Fig 1, S3 Table).

Genetic kinship calculation, phylogenetic tree construction and principal component analysis

Based on the 208,993 high quality SNPs, 87,571 pairwise calculations for all 419 rice landraces were carried out. Of these, the number of pairs reached up to 45,709, which the genetic relationship coefficients <0.05. The phylogenetic tree was clustered two mainly panels in accordance with the *indica* and *japonica* subpopulations. The first two principal components explained 5.64% and 3.92% of the genetic variation respectively (Fig 2).

Analysis of population sturcture

Based on the error rate of 5-fold cross-validation, the ancestor number was confirmed to 6 for all the 419 rice landraces. The six subpanels respectively contain 330 *indica* varieties, 78 *japonica* varities (Fig 3).

GWAS mapping in rice landraces

The whole panel composed of 419 rice landraces, *indica* subpanel (330 rice landraces) and *japonica* subpanel (78 rice landraces) were utilized for GWAS respectively in order to avoid population structure noise. Only one QTL was detected for GT, but in both the whole population and *indica* subpopulation. Similarly, only one association with GC was obtained, also both the whole population and *indica* subpopulation. Interesting, the two QTLs were both located on chromosome 6 (Fig 4, Fig 5). For the pericarp color trait, 8 QTLs were detected in the whole panel on chromosome1, 3,4,7, 8, 10 and 11, and 7 QTLs in the *indica* subpanel on chromosome 3, 4, 7, 8, 10 and 11 (Fig 6).

GWAS on Gelatinization Temperature

Gelatinization temperature is one of the most important indexes to evaluate the cooking and eating quality of rice. The association analysis for GT was conducted for the whole panel, *indica* subpanel and *japonica* subpanel successively. In 2014, for the whole panel of 419 rice accessions, GWAS detected a total of 48 GT related SNPs , the 26 SNPs of which can also be detected in *indica* subpanel, and the significant associated SNPs distributed on the 1807797 bp-7174281 bp of chromosome 6 (Fig 4, S4 Table). In the whole and *indica* panels, the most significant associated SNPs with GT were Chr6_6733351 (P=2.04×10E-15) and Chr6_6740370 (1.39×10E-13). In 2015, 8 GT related SNPs were identified in the whole panel, of which 4 SNPs can also be detected in *indica* panel, and the positions of SNPs ranged from 6740370 bp to 6927719 bp of chromosome 6 (Fig 4, S4 Table). However, no significant SNPs were detected for the *japonica* subpanel in the both 2014 and 2015 year. In the whole panel and *indica* subpanel, the most significant associated SNPs were both Chr6_6879531 (2.23×10E-11, 1.14×10E-9). Chr6_6733351 and Chr6_6740370 were

located in 15kb and 8 kbupstream of the *ALK* gene (*LOC_Os06g12450*), respectively (Table 1).

The *ALK* gene encodes soluble starch synthase a (*SSIIa*) and is responsible for GT of rice [7]. Some researchers have analyzed the SNP in *ALK* gene, and considered that SNP variation is the main factor of gelatinization temperature change [41-43]. We also found a SNP Chr6_1807797 ($P=2.53\times10^{-8}$) which was significantly associated with GT located in the *Wx* gene region in 2014, which further verified the correlation between the *ALK* gene and the *Wx* gene (Fig 4).

GWAS on Gel Consistency

Gel consistency is a complex quality trait, and views about the genetic basis of GC were not consistent between researchers (http://www.gramene.org/). Gel consistency is inversely related to amylose content, we performed Pearson correlation analysis for gel consistency and amylose content trait, and reached the same conclusion (S5 Table). Many researchers believed that the *Wx* gene located on the Chromosome 6 of rice is the major gene controlling the gel consistency [44-46] , and some other GC-related QTLs located on chromosome 1, 2, 3, 6, and 7 were also detected [47,48]. In 2014, in the whole panel and *indica* subpanel, GWAS detected a total of 28 and 24 significantly associated SNPs respectively, the QTL located in the interval of 1607061 bp-1958767 bp of chromosome 6, and the most significant associated SNP were both Chr6_1797551 (P=1.03×10E-17; P=3.62×10E-16) for both whole and *indica* panels (Fig 5, S6 Table). In 2015, 13 and 4 associated SNPs were confirmed for the whole and *indica* panels respectively. The associated SNPs sited in 1661801 bp-1822395 bp of chromosome 6 (Fig 5, S6 Table). However, no SNPs were detected for the *japonica* subpanel in both 2014 and 2015 year. For the whole and *indica* panels, the most significant associated SNPs were Chr6_1807797 (P=9.48×10E-12) and Chr6_1754453 (P=1.86×10E-9) respectively. The Chr6_1754453, Chr6_1797551 and Chr6_1807797 were located 11.2kb upstream, 26.9kb downstream and 37.2kb downstream of the *Wx* gene (*LOC_Os06g04200*) respectively (Table 1).

GWAS on Pericarp Color

Proanthocyanidins and anthocyanins are accumulated in the red pericarp rice and black pericarp rice respectively. The anthocyanin biosynthesis pathway includes multiple structural genes, such as *CHS, CHI* and *DFR*. A large number of studies have shown that these structural gene expressions have different degrees of synergy, which are directly controlled by MBW protein complexes formed by MYB, bHLH and WD40 transcription factors. The synthesis of anthocyanin in most plants is regulated by MBW through binding to the promoter of structural genes.

For the whole and *indica* panels, 763 and 99 significantly associated SNPs were identified respectively, however significant SNPs were not detected for *japonica* panel (Fig 6, S7 Table). For chromosome 1 of rice, 25 significantly associated SNPs were detected in the whole panel, but no SNPs in *indica* subpanel. The most significant SNP was Chr1_22408336 (P=6.80×10E-14), located in the 2.97 Mb

downstream of proanthocyanidins biosynthesis gene *Rd* [49] (*LOC_Os01g44260*). For chromosome 3, 647 and 27 significant associated SNPs were detected for the whole and *indica* pane respectively, which sited in two clearly divided QTLs. In the 17126203 bp-24432074 bp, the most significantly associated loci were Chr3_20743207 (P=1.50×10E-27) and Chr3_17963359 (P=7.88×10E-16) for the whole and *indica* panels respectively. The *LOC_Os03g38210* and *LOC_Os03g31230*, which both encode the MYB family transcription factors, were located in the 459.6 kb downstream of Chr3_20743207 and 179.2kb upstream of Chr3_17963359 respectively (Table 1). In the 31984778 bp-34408499 bp of chromosome 3, for the whole and *indica* panels, the most significantly associated SNPs were Chr3_32304963 (P=1.07×10E-15) and Chr3_31521682 (P=2.89×10E-9); the Chr3_32304963 sited in the 350.3kb downstream of gene *LOC_Os03g56090* encoding MYB family transcription factor; Chr3_31521682 located in 94.8kb downstream of *LOC_Os03g55220* and 86.9kb downstream of *LOC_Os03g55550*, which both encode bHelix-loop-helix transcription factor (Table 1). For chromosome 4, 11 and 5 significantly associated SNPs were found between whole and *indica* panels, and the most significant SNPs were both Chr4_26803164 (P=1.07×10-E12, P=6.35×10E-12) (Table 1). *Kala4* was located in the 1.11 Mb downstream of Chr4_26803164, which encodes a bHLH transcription factor, provided structural rearrangements of its promoter region, then resulting in its ectopic expression, thus giving the trait of black pericarp in rice [50]. For chromosome 7, 12 significant SNPs were detected in the whole panel, and 61 significant SNPs in *indica* subpanel. The most significant SNP was Chr6_6069266 (2.52×10E-10; 1.72×10E-18), located in the *Rc* gene region (Table 1), which encodes a bHLH motif contained protein that participates in the synthesis of proanthocyanidins of pericarp, and the 14-bp deletion of seventh exon of Rc leads to the Rc mutation to *rc* [51]. For chromosome 8, 17 significantly associated SNPs for the whole panel and 2 SNPs for *indica* subpanel were detected. For the whole panel, the gene *LOC_Os08g21660* encoding a WD domain, G-beta repeat domain contained protein located in 59.9kb upstream of Chr8_12968543 (P=3.96×10E-10) (Table 1), which also located in the QTL ranging from 938782 bp to 20998896 bp of *indica* subpanel. For chromosome 10, 37 significant SNPs for the whole panel and 3 significant SNPs for *indica* panel were found, and there was no identical SNPs between them. Forthe whole panel, *LOC_Os10g30690* and *LOC_Os10g30719* (Table 1), which encode MYB family transcription factor, were located in 148.5kb and 162.3kb downstream of the most significant SNP Chr10_15835327 (P=3.25×10E-14) respectively. For Chr11, 14 and 1 significant SNPs were detected for the whole and *indica* panels respectively , but no candidate genes related to the pericarp color were found.

Discussion

SNP markers obtained by SLAF-Seq technology

Rice is the most important staple crop worldwide, and feeding a fast growing population, so it is emergent to identify genes related to agronomically important

traits.The association analysis is powerful to identify phenotypic variance related nucleotide polymorphisms [52,53]. In this study, in order to dissect the genetic basis of gelatinization temperature, gel consistency and pericarp color, we performed GWAS based on SLAF-seq technology. A total of 67,665 SLAF tags were obtained with an average sequencing depth of 8.75×. Ultimately, 56,768 polymorphic SLAF tags were retained when aligned to the reference genome, which polymorphic ratio reached up to 83.89%. A total of 211,818 SNPs were identified and the number of SNPs with MAF > 0.05 and integrity > 0.5 was 208,993, which then were used for association analysis. Wang et al. (2016) [68] have documented that low-coverage whole-genome sequencing is an effective strategy for genome-wide association studies in rice. Our research also confirms the conclusion that the SLAF-technology can effectively and accurately identify the associated genes, though the obtained genome information is less than information, which is obtained by whole genome resequencing.

QTL comparison among different panels

In this study, in order to eliminate the noise of population structure, we conducted association analysis for the whole panel, *indica* subpanel and *japonica* subpanel respectively. The GWAS results in 2014 and 2015showed that the QTLs of GT, GC and PC were relatively stable in different environmental conditions. For gelatinization temperature, one quantitative trait locus on chromosome 6 was detected in the whole panel and *indica* subpanel, which overlapped with *ALK* gene (Table 1), a confirmed major gene for GT [7]. One association with gel consistency was located on chromosome 6 in the whole panel and *indica* subpanel, where *Wx* (*LOC_Os06g04200*) located (Table 1, http://rice.plantbiology.msu.edu/). For the pericarp color trait, eight QTLs were detected in the whole panel on chromosome 1, 3, 4, 7, 8, 10 and 11, and seven QTLs in the *indica* subpanel on chromosome 3, 4, 7, 8, 10 and 11 (Table 1). For the three traits, no QTL was detected in *japonica* subpanel，probably because of the polymorphism repartition between the subpanel [54], or small population size of *japonica* subpanel [25].

Analysis of candidate genes for GT, GC and PC

Core collection is a subset of the germplasm resources, representative of the genetic diversity and geographical distribution of the entire population with the minimum number of genetic resources [55]. Guangxi province of China is likely to be the origin of cultivated rice [2], which possesses a large number of rice germplasms and genetic resources.

In this study, a GWAS for GT, GC and PC of 419 rice landraces core germplasms from guangxi was performed, and it was concluded that *ALK* is the major effect gene for GT [56,57] and *Wx* has a minor effect on gelatinization temperature (2009) [58]. GC is an eating and cooking quality related trait with complicated genetic basis. So far, more than 20 QTLs for gel consistency have been detected (http://www.gramene.org/) in rice, which located on chromosome1, 2, 4, 7, 11.

Many researchers asserted that *Wx* is the major gene for GC [7,45,46,56]. Swamy et al. (2012) [59] identified 6 QTLs for GC located on chromosome 1, 2, 4, 11 in BC2F2 population. Based on RAD-seqtechnology, Peng et al. (2016) [48] confirmed a *Wx* linkaged QTL for GC on chromosome 6 in BC1F5 population derived from crosses between YVB × V20B. Li et al. (2004) [47] found two QTLs for GC on chromosome 2, 7 through RFLP and SSR marker in BC3F1 population obtained from V20A (*O. sativa* L.) ×103544 (*O. glaberrima* S.). In this study, the major effort QTL for GC located on chromosome 6, which overlapped *Wx* gene.

The accumulation of proanthocyanidins and anthocyanins in the testa leads to red and black pericarp respectively. The anthocyanin metabolism pathway in *Zea mays*, *Antirrhinum majus* and *Arabidopsis thaliana* has been explored to some extent [60], which is not yet fully understood in rice. So far, some structural genes involved in anthocyanin biosynthesis has been identified, such as *OsCHS*1 [61], *OsCHS2* [62], *OsCHI* [63], *OsF3H* [64], *OsF3'H* [62], *OsDFR*(*Rd*) [49] and *OsANS* [62]. However, the regulatory genes, MYB, bHLH and WD40 are not fully reported. Furukawa et al. cloned *Rc* gene located on chromosome 7, which encodes bHLH domain contained transcription factor and participates in the proanthocyanidin synthesis [51], which is consistent with our result. Oikawa et al. (2015) [50] identified *Kala4* gene controlling anthocyanin biosynthesis, identical to *OSB2* gene confirmed by Sakamoto et al. (2001) [65] in rice. Chin et al.(2016) [66] have showed that the *OsC1* gene of chromosome 6 is related to the purple sheath in rice, encoding a MYB transcription factor.

In this study, for the whole panel, a pericarp color-related QTL was detected, where the most significant SNP Chr1_22408336 was adjacent to the *Rd* [49] participating in the proanthocyanidins biosynthesis. The *LOC_Os03g38210* , *LOC_Os03g31230* and *LOC_Os03g56090* all encode a MYB transcription factor, which are different locus of *Kala3* (*Os03g0410000*) participating in anthocyanin biosynthesis for black rice [67].

Both *LOC_Os03g55220* and *LOC_Os03g55550* encode bHLH motif contained transcription factor. The anthocyanin biosynthesis related key gene *Kala4* locates in the significant QTL of chromosome 4 [50]. The *LOC_Os08g21660* gene locates in 54.35kb upstream of Chr8_12968543, and encodes a WD domain and G-beta repeat domain contained protein. So far, the WD40 domain contained gene has not been identified in rice, which is involved in the anthocyanin biosynthesis [50]. There are two genes(*LOC_Os10g30690* and *LOC_Os10g30719)* located in 148.5 kb and 162.3 kb downstream of most significant SNP of Chromosome 10 respectively, both encoding MYB transcription factor, which are not overlapped the gene *qPc10* (*Os10g0536400*) identified by Wang et al. (2016) [68].

With the improvement of people's living standard, the rice of high quality is more and more needed. The discovery and utilization of excellent germplasm can accelerate rice breeding. Based on SLAF-seq technology, GWAS for gelatinization temperature, gel consistency and pericarp color of 419 rice core collections from Guangxi was

conducted, and associated genes, especially the anthocyanin synthesis related genes on chromosome 3,8,10 and 11 were reported for the first time. This study shed light on the genetic analysis for important agricultural trait of rice and beneficial to plant breeder.

Acknowledgments

This study was financially supported by National key R & D projects (2016YFD0100101-03), Guangxi's Ministry of Science and Technology (AB16380117), Guangxi Natural Science Foundation of China (2015GXNSFAA139054) and Guangxi Academy of Agricultural Sciences (2015YT15; 2016JM09).

Author Contributions

Xinghai Yang designed, performed the experiment and whrote the manuscript, Xiuzhong Xia, Yu Zeng, Baoxuan Nong, Zongqiong Zhang performed the experiment, Yanyan Wu, Faqian Xiong, Yuexiong Zhang, Haifu Liang, Guofu Deng collected and analyzed data, Li Danting designed and revised the manuscript. All authors reviewed and approved this submission.

Reference:

1. Matsuoka M, Sakamoto T. Identifying and exploiting grain yield genes in rice. Curr Opin Plant Biol. 2008; 11: 209-214. https://doi.org/10.1016/j.pbi.2008.01.009 PMID: 18343712 .
2. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012; 490: 497-501. https://doi.org/10.1038/nature11532 PMID: 23034647 .
3. Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, et al. Control of grain size, shape and quality by *OsSPL16* in rice. Nat Genet. 2012; 44: 950-954. https://doi.org/10.1038/ng.2327 PMID: 22729225.
4. Li Y, Fan C, Xing Y, Jiang Y, Luo L, Sun L, et al. Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. Nat Genet. 2011; 43: 1266-1269. https://doi.org/10.1038/ng.977 PMID: 22019783 .
5. Song XJ, Huang W, Shi M, Zhu MZ, Lin HX. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nat Genet. 2007; 39: 623-630. https://doi.org/10.1038/ng2014 PMID: 17417637.
6. Jiao Y, Wang Y, Xue D, Wang J, Yan M, Liu G, et al. Regulation of *OsSPL14* by OsmiR156 defines ideal plant architecture in rice. Nat Genet. 2010; 42: 541-544. https://doi.org/ 10.1038/ng.591 PMID: 20495565.
7. Gao Z, Zeng D, Cui X, Zhou Y, Yan M, Huang D, et al. Map-based cloning of the *ALK* gene, which controls the gelatinization temperature of rice. Sci China C Life Sci. 2003; 46: 661-668. https://doi.org/ 10.1360/03yc0099 PMID: 18758723 .
8. Wang Y, Ren Y, Liu X, Jiang L, Chen L, Han X, et al. *OsRab5* a regulates endomembrane organization and storage protein trafficking in rice endosperm cells. Plant J. 2010; 64: 812-824.

https://doi.org/10.1111/j.1365-313X.2010.04370.x PMID: 21105928.

9. Li Y, Fan C, Xing Y, Yun P, Luo L, Yan B, et al. *Chalk5* encodes a vacuolar H+-translocating pyrophosphatase influencing grain chalkiness in rice. Nat Genet. 2014; 46: 398-404. https://doi.org/10.1038/ng.2923 PMID: 24633159 .
10. Wang ZX, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, et al. The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. Plant J. 1999; 19: 55-64. PMID: 10417726.
11. Sun X, Cao Y, Yang Z, Xu C, Li X, Wang S, et al. *Xa26*, a gene conferring resistance to Xanthomonas oryzae pv. oryzae in rice, encodes an LRR receptor kinase-like protein. Plant J. 2004; 37: 517-527. PMID: 14756760.
12. Ji H, Kim SR, Kim YH, Suh JP, Park HM, Sreenivasulu N, et al. Map-based Cloning and Characterization of the *BPH18* Gene from Wild Rice Conferring Resistance to Brown Planthopper (BPH) Insect Pest. Sci Rep. 2016; 6: 34376. https://doi.org/10.1038/srep34376 PMID: 27682162 .
13. Salvi s TR. To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci. 2005; 10: 297-304. https://doi.org/10.1016/j.tplants.2005.04.008 PMID: 15949764 .
14. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010; 42: 961-967. https://doi.org/10.1038/ng.695 PMID: 20972439.
15. Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet. 2011; 44: 32-39. https://doi.org/10.1038/ng.1018 PMID: 22138690 .
16. Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet. 2014; 46: 714-721. https://doi.org/10.1038/ng.3007 PMID: 24908251.
17. Si L, Chen J, Huang X, Gong H, Luo J, Hou Q, et al. *OsSPL13* controls grain size in cultivated rice. Nat Genet. 2016; 48: 447-456. https://doi.org/10.1038/ng.3518 PMID: 26950093.
18. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016; 48: 927-934. https://doi.org/10.1038/ng.3596 PMID: 27322545 .
19. Meyer RS, Jy C, Sanches M, Plessis A, Flowers JM, Amas J, et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat Genet. 2016; 48: 1083-1088. https://doi.org/10.1038/ng.3633 PMID: 27500524.
20. Li LF, Li YL, Jia Y, Caicedo AL, Olsen KM. Signatures of adaptation in the weedy rice genome. Nat Genet. 2017; 49: 811-814. https://doi.org/10.1038/ng.3825 PMID: 28369039.

21. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 2013; 45: 43-50. https://doi.org/10.1038/ng.2484 PMID: 23242369 .

22. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A. 2013; 110: 453-458. https://doi.org/10.1073/pnas.1215985110 PMID: 23267105 .

23. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015; 33: 408-414. https://doi.org/ 10.1038/nbt.3096 PMID: 25643055 .

24. Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, et al. Genomic analyses provide insights into the history of tomato breeding. Nat Genet. 2014; 46: 1220-1226. https://doi.org/10.1038/ng.3117 PMID: 25305757 .

25.Farlow AA. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013; 9: 29. https://doi.org/10.1186/1746-4811-9-29 PMID: 23876160 .

26. Platt A, Vilhjálmsson BJ, Nordborg M. Conditions under which genome-wide association studies will be positively misleading. Genetics. 2010; 186: 1045-1052. https://doi.org/10.1534/genetics.110.121665 PMID: 20813880.

27. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 2013; 45: 43-50. https://doi.org/10.1038/ng.2484 PMID: 23242369 .

28. Huang x HB. Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol. 2014; 65: 531-551. https://doi.org/10.1146/annurev-arplant-050213-035715 PMID: 24274033.

29. Little RR, Hiller GB, Son E. Differential effect of dilute alkali on 25 varieties of milled white rice. Cereal Chem. 1958; 35: 111-126.

30. Cagampang GB, Perez CM, Bo JO. A gel consistency test for eating quality of rice. J Sci Food Agric. 1973; 24: 1589-1594. PMID: 4771843.

31. Murray mg TW. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 1980; 8: 432-515. PMID: 7433111.

32. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002; 12: 656-664. https://doi.org/10.1101/gr.229202. Article published online before March 2002 PMID: 11932250 .

33.Li h DR. Fast and accurate short read alignment with Burrows-Wheeler transform . Bioinformatics. 2009; 25: 1754-1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168.

34. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297-1303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199 .

35. Li h HR, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078-2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943 .

36. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011; 28: 2731-2739. https://doi.org/10.1093/molbev/msr121 PMID: 21546353.

37. Hardy O, Vekemans X. SPAGeDi a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes. 2002; 2: 618-620. https://doi.org/10.1046/j.1471-8278.2002.00305.x .

38. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19: 1655-1664. https://doi.org/10.1101/gr.094052.109 PMID: 19648217 .

39. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012; 28: 2397-2399. https://doi.org/10.1093/bioinformatics/bts444 PMID: 22796960.

40. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007; 23: 2265-2633. https://doi.org/10.1093/bioinformatics/btm308 PMID: 17586829.

41. Umemoto T, Aoki N, Lin H, Nakamura Y, Inouchi N, Sato Y, et al. Natural variation in rice starch synthase IIa affects enzyme and starch properties. Funct Plant Biol. 2004; 31: 671-684. https://doi.org/10.1071/FP04009.

42. Nakamura Y, Francisco PJ, Hosaka Y, Sato A, Sawada T, Kubo A, et al. Essential amino acids of starch synthase IIa differentiate amylopectin structure and starch quality between *japonica* and *indica* rice varieties. Plant Mol Biol. 2005; 58: 213-227. https://doi.org/10.1007/s11103-005-6507-2 PMID: 16027975 .

43. Bao JS, Sun M, Corke H. Analysis of the genetic behavior of some starch properties in *indica* rice (*Oryza sativa* L.): thermal properties, gel texture, swelling volume. Theor Appl Genet. 2002; 104: 408-413. https://doi.org/10.1007/s001220100688 PMID: 12582713 .

44. Su Y, Rao Y, Hu S, Yang Y, Gao Z, Zhang G, et al. Map-based cloning proves *qGC-6*, a major QTL for gel consistency of *japonica*/*indica* cross, responds by Waxy in rice (*Oryza sativa* L.). Theor Appl Genet. 2011; 123: 859-867. https://doi.org/10.1007/s00122-011-1632-6 PMID: 21698394.

45. Tran N, Daygon VD, Resurreccion AP, Cuevas RP, Corpuz HM, Fitzgerald MA. A single nucleotide polymorphism in the *Waxy* gene explains a significant component of gel consistency. Theor Appl Genet. 2011; 123: 519-525. https://doi.org/10.1007/s00122-011-1604-x PMID: 21562821 .

46. Gao Z, Zeng D, Cheng F, Tian Z, Guo L, Su Y, et al. *ALK*, the key gene for gelatinization temperature, is a modifier gene for gel consistency in rice. J Integr Plant Biol. 2011; 53: 756-765. https://doi.org/10.1111/j.1744-7909.2011.01065.x

PMID: 21711449.

47. Li J, Xiao J, Grandillo S, Jiang L, Wan Y, Deng Q, et al. QTL detection for rice grain quality traits using an interspecific backcross population derived from cultivated Asian (*O. sativa* L.) and African (*O. glaberrima* S.) rice. Genome. 2004; 47: 697-704. https://doi.org/10.1139/g04-029 PMID: 15284874 .
48. Peng Y, Hu Y, Mao B, Xiang H, Shao Y, Pan Y, et al. Genetic analysis for rice grain quality traits in the YVB stable variant line using RAD-seq. Mol Genet Genomics. 2016; 291: 297-307. https://doi.org/10.1007/s00438-015-1104-9 PMID: 26334612 .
49. Furukawa T, Maekawa M, Oki T, Suda I, Iida S, Shimada H, et al. The *Rc* and *Rd* genes are involved in proanthocyanidin synthesis in rice pericarp. Plant J. 2007; 49: 91-102. https://doi.org/10.1111/j.1365-313X.2006.02958.x PMID: 17163879.
50. Oikawa T, Maeda H, Oguchi T, Yamaguchi T, Tanabe N, Ebana K, et al. The Birth of a Black Rice Gene and Its Local Spread by Introgression. Plant Cell. 2015; 27: 2401-2414. https://doi.org/10.1105/tpc.15.00310 PMID:26362607.
51. Sweeney MT, Thomson MJ, Pfeil BE, Mccouch S. Caught Red-Handed: *Rc* Encodes a Basic Helix-Loop-Helix Protein Conditioning Red Pericarp in Rice. Plant Cell. 2006; 18: 283-294. https://doi.org/10.1105/tpc.105.038430 PMID: 16399804.
52. Hamblin MT, Buckler ES, Jannink JL. Population genetics of genomics-based crop improvement methods. Trends Genet. 2011; 27: 98-106. https://doi.org/10.1016/j.tig.2010.12.003 PMID: 21227531 .
53. Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ, et al. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol. 2015; 24: 110-118. https://doi.org/10.1016/j.pbi.2015.02.010 PMID: 25795170 .
54. Phung NT, Mai CD, Hoang GT, Truong HT, Lavarenne J, Gonin M, et al. Genome-wide association mapping for root traits in a panel of rice accessions from Vietnam. BMC Plant Biol. 2016; 16: 64. https://doi.org/10.1186/s12870-016-0747-y PMID: 26964867 .
55.Brown A. Core collections: a practical approach to genetic resources management. Genome. 1989; 21: 818-824. https://doi.org/doi.org/10.1139/g89-144.
56. Septiningsih EM, Trijatmiko KR, Moeljopawiro S, Mccouch SR. Identification of quantitative trait loci for grain quality in an advanced backcross population derived from the *Oryza sativa* variety IR64 and the wild relative O. rufipogon. Theor Appl Genet. 2003; 107: 1433-1441. https://doi.org/10.1007/s00122-003-1376-z PMID: 14513216 .
57. Yf T, Li JX, Yu SB, Xing YZ, Xu CG, Zhang Q. The three important traits for cooking and eating quality of rice grains are controlled by a single locus in an elite rice hybrid, Shanyou 63. Theor Appl Genet. 1999; 99: 642-648. https://doi.org/10.1007/s001220051279 PMID: 22665200.
58. Tian Z, Qian Q, Liu Q, Yan M, Liu X, Yan C, et al. Allelic diversities in rice

565 starch biosynthesis lead to a diverse array of rice eating and cooking qualities.
566 Proc Natl Acad Sci U S A. 2009; 106: 21760-22176.
567 https://doi.org/10.1073/pnas.0912396106 PMID: 20018713 .
568 59. Swamy BP, Kaladhar K, Shobha RN, Prasad GS, Viraktamath BC, Reddy GA, et
569 al. QTL analysis for grain quality traits in 2 BC2F2 populations derived from
570 crosses between *Oryza sativa* cv Swarna and 2 accessions of *O. nivara*. J Hered.
571 2012; 103: 442-452. https://doi.org/10.1093/jhered/esr145 PMID: 22312119.
572 60. Zx zhu YL. Plant color mutants and the anthocy pathway. Chinese Bulletin of
573 Botany (Chinese). 2016; 51: 107-119. https://doi.org/10.11983/CBB15059.
574 61. Reddy AR, Scheffler B, Madhuri G, Srivastava MN, Kumar A, Sathyanarayanan
575 PV, et al. Chalcone synthase in rice (*Oryza sativa* L.): detection of the CHS
576 protein in seedlings and molecular mapping of the chs locus. Plant Mol Biol. 1996;
577 32: 735-743. https://doi.org/ 10.1007/BF00020214 PMID: 8980525.
578 62. Shih CH, Chu H, Tang LK, Sakamoto W, Maekawa M, Chu IK, et al. Functional
579 characterization of key structural genes in rice flavonoid biosynthesis. Planta.
580 2008; 228: 1043-1054. https://doi.org/10.1007/s00425-008-0806-1 PMID:
581 18726614.
582 63. Druka A, Kudrna D, Rostoks N, Brueggeman R, Von wettstein D, Kleinhofs A.
583 Chalcone isomerase gene from rice (*Oryza sativa*) and barley (Hordeum vulgare):
584 physical, genetic and mutation mapping. Gene. 2002; 302: 171-178.
585 https://doi.org/10.1016/S0378-1119(02)01105-8 PMID: 12527208.
586 64. Kim JH, Lee YJ, Kim BG, Lim Y, Ahn JH. Flavanone 3beta-hydroxylases from
587 rice: key enzymes for favonol and anthocyanin biosynthesis. Mol Cells. 2008; 25:
588 312-316. PMID: 18413994.
589 65. Sakamoto W, Ohmori T, Kageyama K, Miyazaki C, Saito A, Murata M, et al. The
590 Purple leaf (*Pl*) locus of rice: the *Pl(w)* allele has a complex organization and
591 includes two genes encoding basic helix-loop-helix proteins involved in
592 anthocyanin biosynthesis. Plant Cell Physiol. 2001; 42: 982-991.
593 https://doi.org/10.1093/pcp/pce128 PMID: 11577193.
594 66. Chin HS, Wu YP, Hour AL, Hong CY, Lin YR. Genetic and Evolutionary
595 Analysis of Purple Leaf Sheath in Rice. Rice (N Y). 2016; 9: 8.
596 https://doi.org/10.1186/s12284-016-0080-y PMID: 26922355.
597 67. Maeda H, Yamaguchi T, Omoteno M, Takarada T, Fujita K, Murata K, et al.
598 Genetic dissection of black grain rice by the development of a near isogenic line.
599 Breed Sci. 2014; 64: 134-141. https://doi.org/10.1270/jsbbs.64.134 PMID:
600 24987299 .
601 68. Wang H, Xu X, Vieira FG, Xiao Y, Li Z, Wang J, et al. The Power of Inbreeding:
602 NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice
603 Domestication. Mol Plant. 2016; 9: 975-985.
604 https://doi.org/10.1016/j.molp.2016.04.018 PMID: 27179918 .
605
606

Table 1 A sbuset of associated loci and candidate genes.

Panel	Trait	Chromosome	Position	P-value	Cadidate gene	Annotation
Full	GT	6	6733351	2.04E-15	*LOC_Os06g12450*	Soluble starch synthase 2-3
Indica	GT	6	6740370	1.39E-13	*LOC_Os06g12450*	Soluble starch synthase 2-3
Full	GC	6	1797551	1.03E-17	*LOC_Os06g04200*	Starch synthase
Indica	GC	6	1797551	3.62E-16	*LOC_Os06g04200*	Starch synthase
Full	GC	6	1807797	9.48E-12	*LOC_Os06g04200*	Starch synthase
Indica	GC	6	1754453	1.86E-09	*LOC_Os06g04200*	Starch synthase
Full	PC	1	22408336	6.8E-14	*LOC_Os01g44260*	Dihydroflavonol-4-reductase
Full	PC	3	20743207	1.5E-27	*LOC_Os03g38210*	MYB family transcription factor
Indica	PC	3	17963359	7.88E-16	*LOC_Os03g31230*	MYB family transcription factor
Full	PC	3	32304963	1.07E-15	*LOC_Os03g56090*	MYB family transcription factor
Indica	PC	3	31521682	2.89E-09	*LOC_Os03g55220*	MYB family transcription factor
Full	PC	4	26803164	1.07E-12	*LOC_Os04g47059*	bHLH transcription factor
Indica	PC	4	26803164	6.35E-11	*LOC_Os04g47059*	bHLH transcription factor
Full	PC	7	6069266	2.52E-10	*LOC_Os07g11020*	bHLH transcription factor regulating proanthocyanidin production in seeds
Indica	PC	7	6069266	1.72E-18	*LOC_Os07g11020*	bHLH transcription factor regulating proanthocyanidin production in seeds
Full	PC	8	12968543	3.96E-10	*LOC_Os08g21660*	WD domain, G-beta repeat domain containing protein
Indica	PC	8	938782	2.02E-11	*LOC_Os08g21660*	WD domain, G-beta repeat domain containing protein
Full	PC	10	15835327	3.25E-14	*LOC_Os10g30690*; *LOC_Os10g30719*	MYB family transcription factor
Indica	PC	10	5131361	2.32E-12	*LOC_Os10g30690*;	MYB family transcription factor

			LOC_Os10g30719	
Full	PC	11	12447381	8.27E-21
Indica	PC	11	27509828	1.19E-08

GT: gelatinization temperature; GC: gel consistency; PC: pericarp color.

613