INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
20–0994/2021/25–2–388–396
DOI: 10.17957/IJAB/15.1679
http://www.fspublishers.org

Full Length Article

Population Structure Analysis and Genome-Wide Association Study of Rice Landraces from Qiandongnan, China, using Specific–Locus Amplified Fragment Sequencing

Yan Li[1], Xiaofang Zeng[1], Guangzheng Li[1,2], Yi Chen Zhao[3], Jianrong Li[1] and Degang Zhao[1,2,4*]

[1]*The Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region (Ministry of Education), Institute of Agro-Bioengineering and College of Life Sciences, Guizhou University, Guizhou 550025, China*

[2]*The State Key Laboratory Breeding Base of Green Pesticide and Agricultural Bioengineering, Guizhou University, Guiyang, Guizhou, China*

[3]*College of Tea Sciences, Guizhou University, Guiyang 550025, China*

[4]*Guiyang station for DUS testing center of New Plant varieties of MOA. P.R. China in Guizhou Academy of Agricultural Sciences, Guiyang 550006, China*

[*]For correspondence: dgzhao@gzu.edu.cn; yli@gzu.edu.cn

Received 08 July 2020; Accepted 31 October 2020; Published 10 January 2021

Abstract

Uncovering the genetic basis of rice landraces has important applications in breeding. In this study, the specific-locus amplified fragment (SLAF) sequencing method was used to analyze the population structure and conduct a genome-wide association study (GWAS) of the agronomic traits of 60 rice species in Southeast Guizhou. We obtained a total of 178,287,776 reads, 314,065 SLAFs, and 571,521 single nucleotide polymorphisms (SNPs). A neighbor-joining phylogenetic tree, admixture proportions, and principal component analyses revealed that the investigated landraces were divided into *japonica* (heterozygosity rate 0.062) and *indica* (heterozygosity rate 0.073) groups. The groupings were consistent with the local classifications of "He" and "Gu" based on the resistance to seed shattering, and the SNPs clustered in the *qSH1* gene. The GWAS of eight agronomic traits revealed that the signal peaks at four locations were closely related to previously reported genes or gene regions. This study demonstrates that the SLAF sequencing method combined with a GWAS may be effective for investigating the evolution of rice and identifying genes regulating complex traits in rice landraces cultivated in relatively isolated regions.

Keywords: GWAS; Qiandongnan rice landrace; SLAF-seq; SNP

Introduction

Rice (*Oryza sativa* L.) is the most widely grown crop, constituting a staple food for over half of the global population (Lu *et al.* 2009; Wang and Li 2011). Asian cultivated rice consists of the subspecies *japonica* and *indica* (Chen *et al.* 1992; Cheng *et al.* 2003). Research on the genetic differentiation of *indica* and *japonica* rice has resulted in the production of valuable germplasms and important information relevant to crossbreeding (Ikehashi 2009; Lu *et al.* 2009; Jyothi *et al.* 2018; Zhang 2020). Comparative studies of wild, *indica*, and *japonica* rice indicate that wild rice is genetically undifferentiated from *indica* and *japonica* rice (Cheng *et al.* 2003; Shen *et al.* 2004; Song *et al.* 2006), and it was proposed that *japonica* rice was first domesticated from wild rice and that the *indica* subspecies was subsequently developed from crosses between *japonica* and local wild rice (Huang *et al.* 2012).

China is one of the main centers of origin and differentiation for Asian cultivated rice, and consequently, harbors abundant genetic rice resources (Zong *et al.* 2007; Fuller *et al.* 2009; Huang *et al.* 2010). Qiandongnan (26.5834° N, 107.9829° E) is a typical mountainous region (altitude: 1000–2000 m) located on the eastern slope of the Yunnan–Guizhou Plateau in China. Qiandongnan rice is cultivated under relatively isolated conditions in accordance with the local specific environment and agricultural practices and is called "He" and "Gu" by the locals. These landraces have been the staple food for the local Miao and Dong populations and are grown in terraced fields on hills and river valleys. These cultivars possess several favorable agronomic traits, including a desirable awn, grain, and hull color, grain shape, plant height, and growth period length. Thus, cultivated landraces in this region represent ideal models for studying the evolution of rice in mountainous terrains. Identifying the genetic basis for these diverse varieties may

provide important insights into the breeding of elite rice varieties for sustainable agricultural production. However, research involving Qiandongnan rice landraces is limited.

A genome-wide association study (GWAS) based on molecular markers, such as single nucleotide polymorphisms (SNPs), copy number variations, and simple sequence repeats, can be used to analyze gene clustering and quantitative trait loci. Many sequenced plants, including *Arabidopsis thaliana* (Togninalli *et al.* 2020), rice (Huang *et al.* 2010, 2012; Li *et al.* 2018; Yuan *et al.* 2020) and wheat (Ain *et al.* 2015; Bellucci*et al.* 2015; Beyer *et al.* 2019), have been extensively analyzed by GWAS.

Specific-locus amplified fragment sequencing (SLAF-seq), which is based on second-generation sequencing technology, can effectively generate genome-wide high-density markers and efficiently simplify the genome (Sun *et al.* 2013). It has been used to develop high-density molecular markers in rice (Jiang *et al.* 2018; Yang *et al.* 2020), peanut (Hu *et al.* 2018), *Thinopyrum ponticum* (Liu *et al.* 2018), cotton (Chen *et al.* 2014), maize (Xia *et al.* 2014), cauliflower (Zhao *et al.* 2016), soybean (Yang *et al.* 2020) and other plants.

In this study, SLAF-seq analysis was conducted on 60 Qiandongnan rice landraces. A GWAS analysis involving eight agronomic traits was then conducted to identify potential loci associated with rice production and improvement. We identified three loci (signals peak) that were closely related to previously reported genes or gene regions (Matsushita *et al.* 2003; Abe *et al.* 2010; Shao *et al.* 2010; Toriba and Hirano 2014; Tetsuo *et al.* 2015; Li *et al.* 2016). Our results provide new details regarding the evolution of rice and the detection of genes responsible for complex traits in Qiandongnan rice landraces.

Materials and Methods

Plant materials and growth conditions

A total of 60 Qiandongnan rice landraces and two *Oryza rufipogon* Griff. wild rice accessions (Supplementary Table S1) were selected from the rice collection maintained at the Crop Germplasm Resources Centre at the Institute of Agro-Bioengineering of Guizhou University, Guizhou, China. We referred to germplasm database records for details regarding phenotypic variations and geographic origins. For each accession, 30–50 seedlings were cultivated under natural field conditions during normal rice-growing seasons at the Experimental and Demonstration Base for Transgenic Plants at the Agricultural Bioengineering Research Institute (Guiyang, China).

For each landrace, 10 plants were randomly selected. The agronomic traits such as plant height, panicle length, and flag leaf length, width, and angle were calculated in the mature period. The flag leaf length, width, and angle were measured on the main tiller. Awn length and grain length and width were measured using a caliper. Pericarp and awn colors were also observed. The growth period was recorded as the number of days from sowing to fully ripened.

DNA isolation

For each landrace, 20–30 rice seedlings were planted in pots at the Agricultural Bioengineering Research Institute. Total genomic DNA of 20-day-old seedling leaves was extracted according to the cetyl trimethylammonium bromide (CTAB) method (Doyle 1990). The isolated DNA was quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific, USA).

SLAF-seq library construction and high-throughput sequencing

The SLAF-seq procedure was performed as described (Sun *et al.* 2013), with minor modifications. Briefly, a pilot SLAF-seq experiment was conducted to establish the optimal yield conditions to prevent SLAF duplication and produce an even distribution of SLAFs for maximum efficiency. Based on the pilot experiment results, the SLAF library was constructed as follows. Genomic DNA was first incubated at 37°C with *Rsa*I [New England Biolabs (NEB), Ipswich, MA, USA], T4 DNA ligase (NEB), ATP (NEB), and *Rsa*I adapter. Restriction enzyme digestion/ligation reactions were heat-inactivated at 65°C and samples were digested with *Eco*RI and *Bfa*I at 37°C. A polymerase chain reaction (PCR) was conducted using a solution consisting of diluted restriction enzyme digestion/ligation samples, dNTP, Taq DNA polymerase (NEB) and *Rsa*I primer containing barcode 1 (forward primer sequence: 5′-AATGATACGGCGACCACCGA-3′; reverse primer sequence: 5′-CAAGCAGAAGACGGCATACG-3′). The following PCR amplification conditions were used: 98°C for 3 min; 18 cycles of 98°C for 10 s, 65°C for 30 s and 72°C for 30 s; 72°C for 5 min. The PCR products were purified using an E.Z.N.A. Cycle Pure Kit (Omega Bio-Tek, Norcross, GA, USA) and pooled. The pooled samples were incubated at 37°C with *Rsa*I, T4 DNA ligase, ATP, and Solexa adapter. Samples were then purified using a Quick Spin column (Qiagen, Hilden, Germany) and analyzed on a 2% agarose gel. Fragments that were 350–450 bp or 500–600 bp (with indices and adaptors) were isolated using a Qiagen QIAquick Gel Extraction Kit (Qiagen). The fragments were then amplified using the Phusion Master Mix (NEB) and Solexa Amplification primer mix (Illumina, San Diego, CA, USA) according to the manufacturers' recommended procedures. After the samples were gel purified, DNA bands (SLAFs) corresponding to 350–450 bp or 500–600 bp were excised and diluted for paired-end sequencing using a HiSeq 2500 sequencing platform (Illumina). Low-quality reads (Q < 20), reads with adaptor sequences, and duplicated reads were eliminated, and the remaining high-quality reads were used for mapping.

SNP calling and quality

The 125-bp raw read pairs were filtered and trimmed so that each mate was at least 80 bp, with a minimum Q_{phred} quality value > 20 over three consecutive bases. We used the Rice Genome Annotation Project reference genome (382 Mb; RGAP v. 7.0; http://rice.plantbiology.msu.edu/index.shtml) (Goff *et al.* 2002). The original reads for each sample were aligned with the reference genome sequence using SOAP software (Li *et al.* 2009a). Reads that were mapped to the same positions were classified into the same SLAF group. If the reads mapped to the reference genome overlapped with two SLAF tags, they were assigned to both groups (Gu *et al.* 2015). The number of SLAF markers per 100 kb was calculated to determine the SLAF marker distribution on each chromosome.

To identify high-quality SNPs, the GATK (McKenna *et al.* 2014) and Samtools (Li *et al.* 2009b) packages were used for SNP calling based on the defined SLAF Markers. The SNPs identified by GATK and Samtools were selected, and those with a genetic integrity ≥ 50% and a minor allele frequency ≥ 5% were used for further analysis. The SLAF markers with SNPs were considered polymorphic SLAF markers and were used to calculate the number of SNPs per 100 kb (*i.e.*, SNP distribution on each chromosome).

Phylogenetic and population genetic analyses

The population structure of the 60 collections was determined with high-quality SNPs using the Admixture program (Alexander *et al.* 2009), which estimates individual ancestry and admixture proportions. The number of clusters (K) was predefined as 2–10, and the best K result was used for the Q matrix in the GWAS.

Neighbor-joining trees and principal component analysis (PCA) plots were used to characterize the population structure of the rice landraces. A phylogenetic tree consisting of 60 local landraces was generated according to neighbor-joining analyses using MEGA5 software (Saitou and Nei 1987; Tamura 2011). The PCA was conducted using cluster analysis software (Shaun *et al.* 2007) to clarify the clustering of the main component. Neighbor-joining trees were produced and PCA analysis was conducted based on SNPs.

Genome-wide association analysis

SNP markers were used to analyze the associations between genes and quantitative traits. The TASSEL program (Bradbury *et al.* 2007) was used to determine the association between each SNP and the corresponding trait, and association graphs were plotted with Admixture and SPAGeDi software (Hardy and Vekemans 2002).

The association between SNP markers and the corresponding trait (Supplementary Table S2) was determined using the simple linear model (GLM compressed) of the TASSEL software (Bradbury *et al.* 2007). As there were both *japonica* and *indica* rice groups in the population, it was not suitable to use the compressed MLM method. We used the following formula: $Y=X\alpha+Q\beta+K\mu+e$, where Q represents the population structure calculated by Admixture, K represents the genetic relationship between samples calculated using SPAGeDi software (Hardy and Vekemans 2002), e represents the residual term, X corresponds to genotype, and y refers to phenotype. The *P*-value for each SNP was calculated. Gene annotations were based on the reference genome of the Rice Genome Annotation Project (382 Mb; RGAP v. 7.0; http://rice.plantbiology.msu.edu/index.shtml) (Goff *et al*. 2002).

Accession numbers

The gene sequence data from this article can be found in the Rice MSU Genome Annotation Release 7 under the following accession numbers: *qSH1* gene (LOC_Os01g62920, Chr1: 36.44 M), *An10* gene (RM237-RM265, Chr7: 28.57 M-Chr7: 36.95 M), *An1* gene (LOC_Os04g28280.1, Chr4: 16.73 M), *pre-awn1*gene (LOC_Os07g35870.1, Chr7: 21.46 M), *Kala4* gene (LOC_Os04g47059.1, chr4: 27.91 M), *pre-anth1* gene (LOC_Os04g47040.1, Chr4: 27.88 M), *pre-anth2* gene (LOC_Os04g47080.1, Chr4: 27.94 M), *DEP2* gene (LOC_Os07g42410, chr7: 25.38 Mb), and *qGL7-2* gene (Indel1-RM21945, chr7: 25.38 M-chr7: 25.62 M).

Results

SLAF-seq

A total of 178,287,776 clean reads were obtained. For each landrace, 2,000,000–3,500,000 clean reads were generated (Supplementary Table S2). The base-calling accuracy (Q score > 20) was 85.98%. Based on the alignment results, 314,065 high-quality SLAFs were obtained, with a 3.20-fold average sequencing depth. The number of SLAFs on each chromosome is shown in Table 1. Among all the chromosomes, chromosome 1 contained the most SLAFs (37,489), while chromosome 9 had the least (19,069). The number of SLAFs per 100 kb was calculated, and the chromosomal distribution of the markers is presented in Fig. 1a. The overall SLAF marker density was 822.1 per Mb (Fig. 1a). This suggested that the SLAF markers were evenly distributed, and that individual SLAF segments were representative of the whole genome. Of the high-quality SLAFs, 165,922 were polymorphic, indicating that the average polymorphism rate was 52.83% (Table 3).

SNP markers

A total of 35,434,302 SNPs were detected with a genetic integrity of 80.67% based on the defined SLAF markers. The identified SNPs were aligned with the reference genome sequence. We obtained 571,521 high-quality SNP

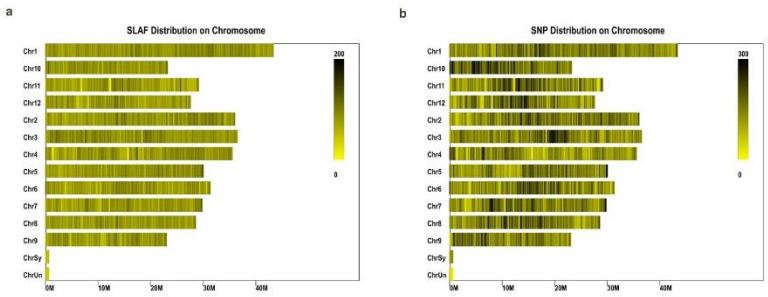

Fig. 1: Distribution maps of SLAF markers and SNPs on rice chromosomes.
(**a**) Distribution map of SLAF markers on rice chromosomes. Each line represents a chromosome. Bar colors correspond to the number of SLAF markers per 100 kb. The bar is black if there are more than 30 SLAFs per 100 kb
(**b**) Distribution map of SNPs on rice chromosomes. Each line represents a chromosome. Bar colors correspond to the number of SNPs per 100 kb. The bar is black if there are more than 300 SNPs per 100 kb

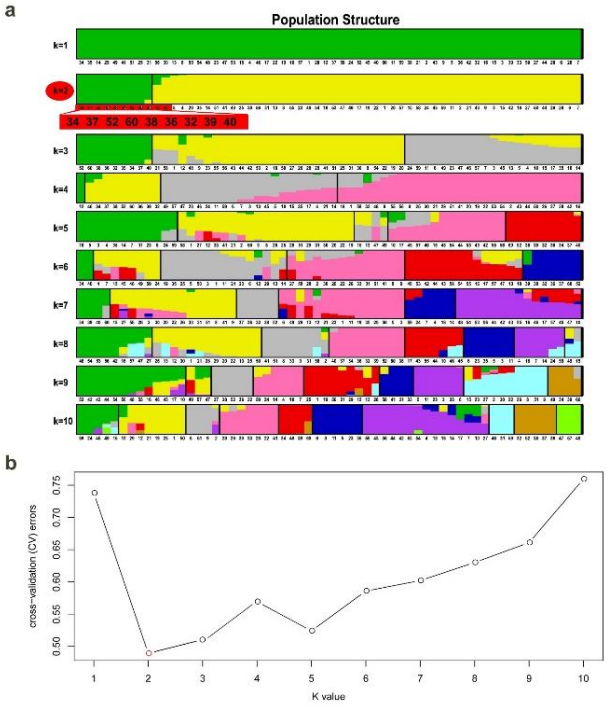

Fig. 2: Cluster map of Qiandongnan rice landraces
(**a**) Analyses of cross-validation errors of clustering with a hypothetical k-value of 1–10. The *x*-axis presents the k-value (1–10), while the *y*-axis presents the cross-validation error values
(**b**) Population structure of 60 landraces with a k-value of 2. Different colors represent individual landraces

markers with an overall density of 1496.1 per Mb. Additionally, the number of SNP markers per 100 kb (Fig. 1b) revealed that the markers were evenly distributed on the chromosomes. In total, 2,302,820 heterozygous SNPs were detected and the heterozygosity rate for all SNPs was 8.06% (Table 2).

Evolutionary structure and analysis of seed shattering

According to the SNPs, all 60 landraces were divided into two populations, *indica* and *japonica* (Fig. 2). Among them,

Table 1: Statistics of SLAF markers on each rice chromosome

Chromosome	SLAF Num.
Chr01	37,489
Chr02	30,897
Chr03	32,621
Chr04	29,761
Chr05	25,796
Chr06	25,673
Chr07	24,458
Chr08	23,459
Chr09	19,069
Chr10	19,765
Chr11	22,615
Chr12	21,809
ChrSy	313
ChrUn	340
Total	314,065

Total is the number of SLAF markers for the entire genome

51 landraces were *japonica* and belonged to the "He" rice group. The smaller *indica* group contained nine landraces, which were present in the "Gu" group (Fig. 2a).

The neighbor-joining algorithm was used to generate the phylogenetic tree. The phylogenetic tree revealed that the 60 landraces were also divided into the same two groups. Fifty-one landraces formed the *japonica* group, while the remaining landraces were included in the *indica* group (Fig. 3a). The *japonica* and *indica* groups corresponded with the "He" and "Gu" groups, respectively. The "Hai nan ye jing" and "Hai nan ye xian" wild rice lines clustered in between the two groups. The "Hai nan ye jing" rice was more closely related to Nipponbare rice, while "Hai nan ye xian" rice was more related to 93-11 rice (Fig. 3a). The PCA clustering results were consistent with those obtained using the neighbor-joining algorithm (Fig. 3b) (Saitou and Nei 1987; Tamura 2011), which supported the classification of Qiandongnan rice landraces into *japonica* or *indica* groups.

The loss of seed shattering is a crucial event during the domestication of cereal crops (Saeko *et al*. 2006). The *qSH1* gene was reported to have an important effect on the grain shattering of rice. In our study, we found that two SNPs (site Chr04: 364,459,91 and Chr04: 364,475,26) were located in the *qSH1* gene, with the A and A bases at Chr04: 364,459,91

Table 2: SNP statistics

SNP Num.	Total Base	Total N	Integrity	Heter Num.	Hete Ratio
571,521	35,434,302	6,848,746	80.67%	2,302,820	8.06%

SNP Num. is the number of total detected SNPs; Total Base is the number of SNPs of all samples; Total N is the number of gene deletion sites; Integrity is the SNPs integrity; Heter Num. is the number of heterozygous SNPs of all samples; Hete Ratio is the heterozygosity rate for all SNPs

Table 3: Polymorphism statistics of the SLAF markers

SLAF Num.	Total Depth	Average Depth	SLAF Poly	Poly Ratio
314,065	62,384,746	3.20	165,922	52.83%

SLAF Num. is the number of all detected SLAF Markers; Total Depth is the number of reads of all SLAF Markers; Average Depth is the average depth of the SLAF markers in each sample; SLAF Poly is the number of polymorphic SLAF markers; Poly Ratio is the percentage of polymorphic SLAF markers

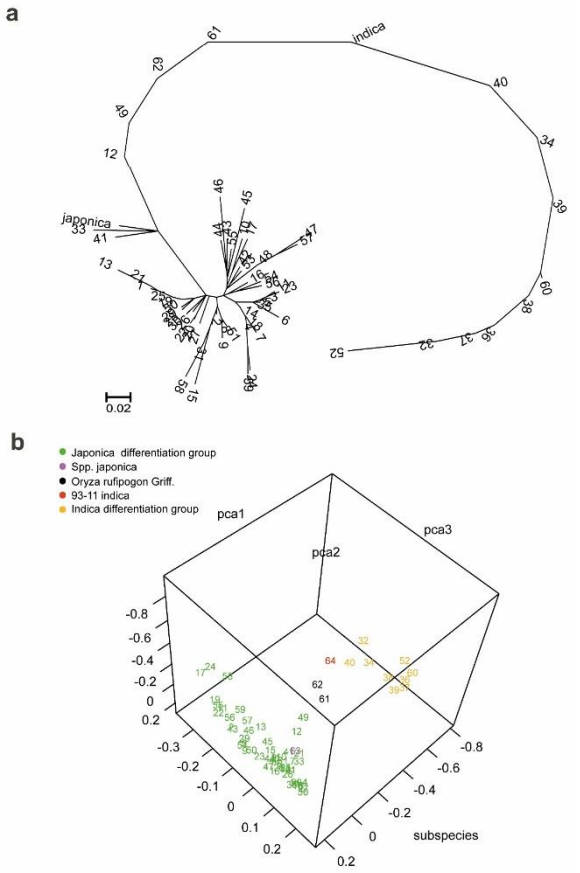

Fig. 3: Phylogenetic tree and PCA of Qiandongnan rice landraces
(**a**) Phylogenetic tree of landraces. Each branch represents a landrace. Samples are also presented in Supplementary Table S1
(**b**) PCA of landraces. The rice landraces were clustered in a 2D graph by PCA. The *x*-axis presents the first principal component and the *y*-axis presents the second principal component. A dot corresponds to a rice sample, while different colors are used to indicate different groups. The samples are also presented in Supplementary Table S1. *indica*: 93-11; *japonica*: Nipponbare

and Chr04: 364,475,26 present in the 51 He landraces (which was consistent with Nipponbare) and both G bases present in the nine Gu landraces (which was consistent with 93-11) (Table 4). Further, the He landraces were consistent with Nipponbare in their shattering character, with the seeds being

Table 4: SNP loci analysis of the qSH1 gene in rice landraces in the Qiandongnan area in Guizhou

Sample No.	SNP locus		Sample No.	SNPs locus	
	Chr1 36445991	Chr1 36447526		Chr1 36445991	Chr1 36447526
1	A	A	33	A	A
2	A	A	34	G	G
3	A	A	35	A	A
4	A	A	36	G	G
5	A	A	37	G	G
6	A	A	38	G	G
7	A	A	39	R	G
8	A	A	40	N	G
9	A	A	41	A	A
10	A	A	42	A	A
11	A	A	43	A	A
12	R	A	44	A	A
13	A	A	45	A	A
14	A	A	46	A	A
15	A	A	47	A	A
16	A	A	48	A	A
17	A	A	49	N	N
18	A	A	50	A	A
19	A	A	51	A	A
20	A	A	52	G	G
21	A	A	53	A	A
22	A	A	54	A	A
23	A	A	55	A	A
24	A	A	56	A	A
25	A	A	57	A	A
26	A	A	58	A	A
27	A	A	59	A	A
28	A	A	60	G	G
29	A	A	61	A	A
30	A	A	62	R	G
31	A	A	63	A	A
32	G	G	64	G	G

Samples 1 to 60 are the local rice landraces, and the samples with "gu" as the last word in their names are marked with bold font. The rest of the samples are "He"
Sample No. 61 is Hai nan ye xian, and sample No. 62 is hai nan ye jing
Sample No. 64 is 93-11 and No. 63 is Nipponbare
R represents A and G, and N is the omitted base in sequencing

resistant to shattering, while the Gu landraces were consistent with 93-11 in that their seeds shattered easily (Saeko *et al.* 2006; Zhang *et al.* 2009).

Analysis of quantitative trait associations

The SNP markers were used to analyze the associations between genes and quantitative traits in the Qiandongnan rice landraces (Hardy and Vekemans 2002; Bradbury *et al.* 2007). Eight genome-wide association graphs were plotted, including the traits of awn length, grain color, grain shape, awn color, flag leaf angle, flag leaf length, grain width, and growth period (Fig. 4 and Supplementary Fig. S1). Here, traits of awn length, grain color, and grain shape were considered as representative examples of the analysis. A total of 2,307 association signals were identified with a P-value < 10^{-6}, and among them we identified 858 strong association signals with a $P < 10^{-8}$ (Table 5). There were significant association signal peaks for awn length on chromosomes 1, 3, 7, and 11 (Fig. 4a, indicated by blue arrows), grain color on chromosomes 3, 4, 8, and 12 (Fig. 4b, indicated by blue

Table 5: Number of association signals in each chromosome

Trait	Marker *P*-value	Num. of association signals												
		Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Chr12	total
Awn length	Marker p (10^{-6})	22	3	12	12	14	15	11	74	4	1	313	7	488
	Marker p (10^{-8})	5	1	2	0	0	0	21	0	0	0	99	0	128
Grain color	Marker p (10^{-6})	161	92	378	172	74	92	93	120	67	80	61	382	1,772
	Marker p (10^{-8})	47	25	237	78	22	26	31	33	30	37	20	130	716
Grain shape	Marker p (10^{-6})	7	4	1	2	3	4	7	12	1	4	1	1	47
	Marker p (10^{-8})	3	1	0	1	2	1	2	1	1	2	0	0	14

Table 6: The regions of the signal peaks

Traits	Signal peak	Region	Marker *P*-value (10^{-6})	Marker *P*-value (10^{-8})	Annotated gene Num.
Awn length	Awn peak 1	Chr1: 34.57–34.66M	13	5	25
	Awn peak 2	Chr3: 27.87–27.97M	5	2	19
	Awn peak 3	Chr7: 21.36–21.61M	71	20	59
	Awn peak 4	Chr11: 21.24–22.97M	303	94	266
Grain color	Grain color peak 1	Chr3: 13.85–22.80M	312	225	1507
	Grain color peak 2	Chr4: 27.38–28.17M	11	9	159
	Grain color peak 3	Chr8: 2.89–3.08M	10	9	32
	Grain color peak 4	Chr12: 3.04–4.43M	46	36	265
	Grain color peak 5	Chr12: 19.40–21.32M	146	53	313
Grain shape	Grain shape peak 1	Chr7: 25.37–25.63M	4	0	53
	Grain shape peak 2	Chr8: 6.45–7.90M	10	0	115

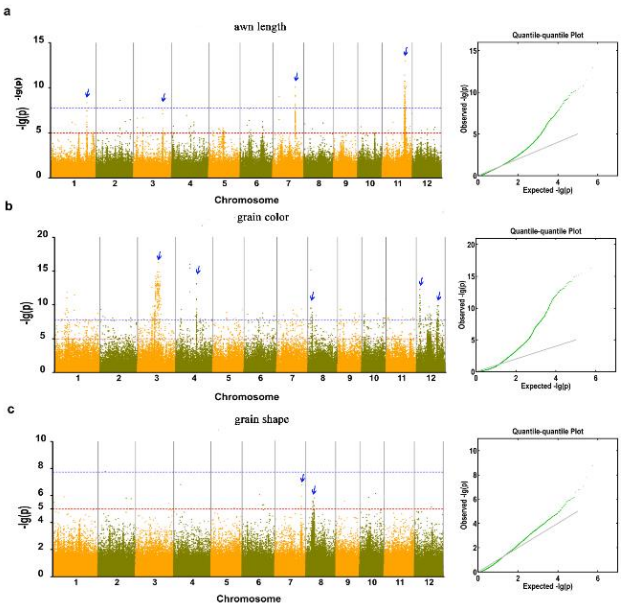

Fig. 4: GWAS of awn length, grain color, and grain shape of the Qiandongnan rice landraces

(**a**) Manhattan plots of the GLM for awn length and quantile-quantile plot of the GLM for awn length. (**b**) Manhattan plots of the GLM for grain color and quantile-quantile plot of the compressed GLM for grain color

(**c**) Manhattan plots of the GLM for grain shape and quantile-quantile plot of the compressed GLM for grain shape

Negative log10-transformed *P*-values from a genome-wide scan are plotted against the positions on each of 12 chromosomes. The yellow horizontal dashed line corresponds to −log10 (*p*) = 5, which is the genome-wide significance threshold. The blue horizontal dashed line corresponds to −log 10 (*p*) = 8, and the blue arrow shows the significant association signal peaks

arrows), and grain shape on chromosomes 7 and 8 (Fig. 4c, indicated by blue arrows). These association signal peak regions are given in Table 6. The gene annotation information was obtained from the Rice Genome Annotation Project 7.0, http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/.

Some of the association signals identified in this study were consistent with the gene or gene location regions reported in other studies. We identified an association signal with rice awn length on chromosome 1 at Chr1: 34.57–34.66 M (Table 6, with 13 associated markers). The results indicated that the region corresponded to the position of the *An10* gene (Matsushita *et al.* 2003), which occurs between two SSR markers (RM237 and RM265) at position 28.57–36.95 Mb on the long arm of chromosome 1 (Matsushita *et al.* 2003) (Fig. 5a and Table 6). It was reported that *An1* encodes a bHLH protein that is essential for rice awn development (Li *et al.* 2016). In the association peaks analysis for awn length traits, we detected a candidate gene for awn length (predicted awn length gene1, *pre-awn1*), annotated as a bHLH protein in a strong association signal peak with 59 associated markers, at chr7: 21.36–21.61 M (Fig. 5b, Table 6).

In terms of grain color, an association signal peak with 11 associated markers was detected at position 27.38–28.17 Mb on chromosome 4 (Table 6). This is consistent with the *Kala4* gene (Tetsuo *et al.* 2015) and another two genes annotated to encode anthocyanin regulatory Lc proteins (predicted anthocyanin regulatory gene 1 and 2, *pre-anth1* and *pre-anth2*), which were predicted to regulate anthocyanin biosynthesis (Fig. 5c).

We also identified an association signal with four markers at the region at position 25.37–25.63 Mb on chromosome 7 (Fig. 5d, Table 6) corresponding to grain shape. This is consistent with a grain shape-related gene *DEP2*, which was reported to regulate grain size (Abe *et al.* 2010). Additionally, this association signal was also close to the region between Indel1 and RM21945 corresponding to the reported location of the grain length-related gene *qGL7-2* (Shao *et al.* 2010).

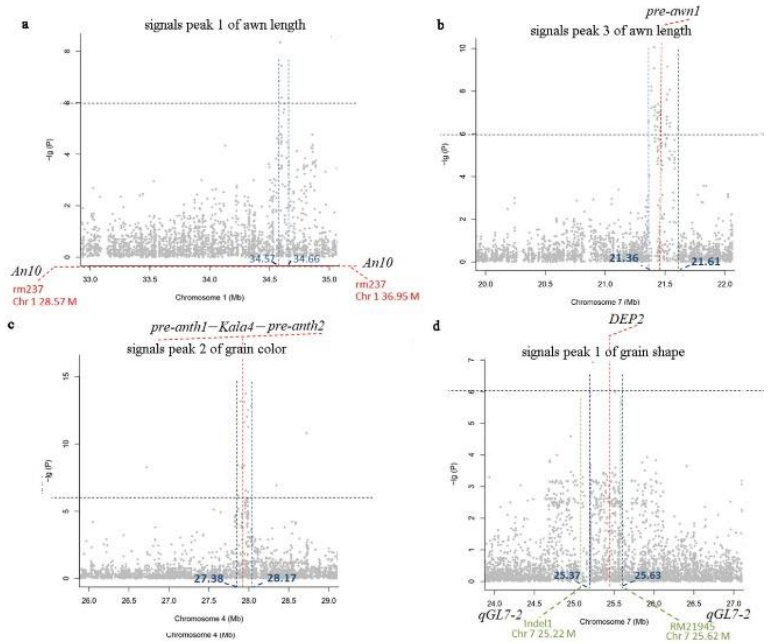

Fig. 5: Regions of the genome showing association signals near previously-identified genes
(**a**) Awn length signal peak 1. (**b**) Awn length signal peak 3. (**c**) Grain color signal peak 2. (**d**) Grain shape signal peak 1
The top of each panel shows a region on each side of the peak SNP, the position of which is indicated by a vertical blue line. Negative log 10-transformed *P*-values from the compressed GLM are plotted on the vertical axis; axis scales are slightly different across panels. Black horizontal dashed lines indicate the genome-wide significance threshold with a *P*-value=6. The bottom of each panel shows a region on each side of the SNP signal peak (blue dashed lines). Previously identified genes are indicated by red dashed lines or green dashed lines

Discussion

SLAF and SNP density is crucial for evolutionary and GWAS analysis. One previous study obtained 144,754 high-quality SLAFs and 30,578 polymorphic SLAFs in soybean, and a total of 8,738 SNPs were selected for gene mapping (Yang *et al.* 2020); another study obtained 374,265 SLAFs, 56,295 polymorphic SLAFs, and 102,025 SNPs in *Ammopiptanthus mongolicus* (Duan *et al.* 2019); 433,679 high-quality SLAFs and 29,075 polymorphic SLAFs were developed in peanut (Hu *et al.* 2018); and in rice, 56,768 SLAFs and 211,818 SNPs were obtained in one study (Yang *et al.* 2018), and 50,330 SLAFs were obtained in another (Li *et al.* 2016). In our study, a total of 178,287,776 reads, 314,065 SLAFs, 165,922 polymorphic SLAFs, and 571,521 evenly-distributed SNPs were developed from 60 Qiandongnan rice landraces, providing a strong foundation for the follow-up analysis.

The Qiandongnan rice landraces were divided into two groups. The *japonica* group contained 51 closely clustered individuals (Fig. 3) with a heterozygosity rate of 0.062. Thus, it can be considered that the sources of these germplasms may be relatively uniform. However, the nine individuals from the *indica* group were not closely clustered (Fig. 3) and had a heterozygosity rate of 0.073, suggesting that the sources of these germplasms were distantly related. This is consistent with the idea that *indica* rice was developed from crosses between *japonica* rice and local wild rice (Huang *et al.* 2012). Additionally, the *japonica* group comprised more members than the *indica* group, which is similar to the current distribution of *japonica* and *indica* rice (Lu *et al.* 2009). Thus, two possibilities exist: the *japonica* and *indica* groups of the Qiandongnan landraces likely evolved locally from their wild rice ancestors during a long-term domestication process and gradually formed the Qiandongnan landrace populations, or alternatively, differentiated landraces were introduced directly to the region. The “Gao qian xiang” and “Gong gui he” landraces were genetically close to the wild rice “Hai nan ye jing” and “Hai nan ye xian”, which clustered between Nipponbare rice and wild rice (Fig. 4). This suggests they may be relatively primitive germplasms. We suspect that the “Gao qian xiang” and “Gong gui he” landraces might be intermediate rice types between wild rice and *japonica* rice.

The results from the admixture groupings, MEGA5 clustering analysis, and cluster PCA enabled the classification of cultivated rice landraces in Guizhou, China, into *japonica* or *indica* groups (Fig. 3 and 4) (Saitou and Nei 1987; Price *et al.* 2006; Shaun *et al.* 2007; Alexander *et al.* 2009; Tamura 2011). These classifications were consistent with the “He” and “Gu” groupings (Table S1) as well as with the SNP sites for the *japonica* and *indica* seed-shattering gene *qSH1* (Saeko *et al.* 2006), which was associated with rice domestication (Table 2). It was previously reported that SNP sites between Nipponbare and 93-11 rice may lead to altered seed shattering (Saeko *et al.*

2006; Zhang *et al.* 2009; Magwa *et al.* 2016; Sheng *et al*. 2019). Our findings suggested that the SNPs in *qSH1* may have important consequences for the domestication (the trait of seed shattering) of Qiandongnan rice landraces. Thus, we suggest that the *qSH1* gene can be used as a marker for investigations of other Qiandongnan rice landraces. Further, we hypothesize that rice growers in the Qiandongnan area classified their rice landraces into "He" and "Gu" based on the resistance to seed shattering over a long period of rice cultivation and the *qSH1* gene may have played an important role in this domestication event.

Kala4 activates the expression of genes upstream of the flavonol biosynthesis pathway, including the genes encoding chalcone synthase and dihydroflavonol 4-reductase. It also induces the expression of downstream genes, such as those for *leucoanthocyanidin reductase* and *leucoanthocyanidin dioxygenase*, to produce a particular pigment. A previous study revealed that the sequence differences from -11 kb to-83 kb upstream of *Kala4* were responsible for rice seed color variability (Tetsuo *et al.* 2015). Interestingly, we discovered that the other two predicted genes *pre-anth1* and *pre-anth2* that regulate anthocyanin were close to *Kala4*. These genes are encoded by a putative anthocyanin regulatory LC protein (Fig. 4b). We hypothesized that these two genes may form a gene cluster with *Kala4* and collectively help regulate rice seed color.

Our GWAS results for awn length and grain shape (Fig. 4) suggest that these traits are regulated by *An10* and *qGL7-2*, respectively (Matsushita *et al.* 2003). These findings may provide relevant information for the discovery of *An10* and *GL7-2* genes. Additionally, some GWAS regions that generated strong association peaks were detected for the first time. No related genes have been reported in these regions, such as Chr11: 21.24–22.97 M, associated with awn length traits (Table 6, with 13 associated markers). We believe these association results will be helpful for the identification of genes responsible for specific traits in rice. However, gene cloning studies are needed to confirm our results.

Conclusion

This study provided a foundation for gene cloning and molecular marker-assisted breeding of these materials and demonstrates the utility of SLAF-seq in a relatively isolated mountainous area.

Acknowledgments

This work was supported by grants from the Genetically Modified Organisms Breeding Major Projects of China [2016ZX08010003], the Natural Science Foundation of Guizhou (20181044), The Young Scholars and Technology Talents Development Project of Guizhou Education Department, the Science and Technology Project of Guizhou Province [20171039], and the Construction Program of Biology First-class Discipline in Guizhou (GNYL[2017] 009).

Author Contributions

Degang Zhao and Yan Li designed the experiment and analyzed the data; Yan Li, Xiaofang Zeng, Guangzheng Li, Yi Chen Zhao and Jianrong Li performed the experiments; Yan Li wrote the manuscript.

Conflict of Interest

The authors have no conflict of interests to declare

References

Abe Y, K Mieda, T Ando, I Kono, M Yano, H Kitano, Y Iwasaki (2010). The small and round seed1 (SRS1/DEP2) gene is involved in the regulation of seed size in rice. *Genes Genet Syst* 85:327–339

Ain QU, A Rasheed, A Anwar, T Mahmood, M Imtiaz, T Mahmood, X Xia, Z He, U Quraishi (2015). Genome-wide association for grain yield under rainfed conditions in historical wheat cultivars from Pakistan. *Front Plant Sci* 6; Article 743

Alexander DH, J Novembre, K Lange (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genomics Res* 19:308–325

Bellucci A, AM Torp, S Bruun, J Magid, SB Andersen, SK Rasmussen (2015). Association mapping in scandinavian winter wheat for yield, plant height, and traits important for second-generation bioethanol production. *Front Plant Sci* 6; Article 1046

Beyer S, S Daba, P Tyagi, H Bockelman, M Mohammadi (2019). Loci and candidate genes controlling root traits in wheat seedlings—a wheat root GWAS. *Funct Integr Genomics* 19:91–107

Bradbury PJ, Z Zhang, DE Kroon, TM Casstevens, Y Ramdoss, ES Buckler (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635

Chen W, J Yao, L Chu, Y Li, XM Guo, YS Zhang, J Jenkins (2014). The development of specific SNP markers for chromosome 14 in cotton using next-generation sequencing. *Plant Breed* 133:256–261

Chen WB, YI Sato, I Nakamura, H Nakai (1992). Indica-japonica differentiation in Chinese rice landraces. *Euphytica* 74:195–201

Cheng CY, M Reiko, T Suguru, F Yoshimichi, O Hisako, O Eiichi (2003). Polyphyletic origin of cultivated rice based on the interspersion pattern of SINEs. *Mol Biol Evol* 20:67–75

Doyle J (1990). Isolation of plant DNA from fresh tissue. *Focus* 12:13–15

Duan YZ, JW Wang, ZY Du, FR Kang (2019). SNP sites developed by Specific Length Amplification Fragment Sequencing (SLAF-seq) and genetic analysis in *Ammopitanthus mongolicus. Bull Bot Res* 38:141–147

Fuller DQ, L Qin, Y Zheng, Z Zhao, XG Chen, Hosoya, L Aoi, GP Sun (2009). The domestication process and domestication rate in rice: Spikelet bases from the Lower Yangtze. *Science* 323:1607–1610

Goff SA, R Darrell, L Tien-Hung (2002). A draft sequence of the rice genome (*Oryza sativa* L. sspp. *japonica*). *Science* 296:92–100

Gu BG, TY Zhou, JH Luo, H Liu, YC Wang, YY Shangguan, JJ Zhu (2015). An-2 encodes a cytokinin synthesis enzyme that regulates awn length and grain production in rice. *Mol Plant* 8:1635–1650

Hardy OJ, X Vekemans (2002). Spagedi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2:618–620

Hu XH, SZ Zhang, HR Miao, FG Cui, Y Shen, WQ Yang, TT Xu, N Chen, XY Chi, ZM Zhang, J Chen (2018). High-density genetic map construction and identification of QTLs controlling oleic and linoleic acid in peanut using SLAF-seq and SSRs. *Sci Rep* 8; Article 5479

Huang XH, N Kurata, XH Wei, ZX Wang, A Wang, Q Zhao, Y Zhao, KY Liu, HY Lu, WJ Li, YL Guo, YQ Lu, CC Zhou, DL Fan, QJ Weng, CR Zhu, T Huang, L Zhang, YC Wang, L Feng, H Furuumi, T Kubo, T Miyabayashi, XP Yuan, Q Xu, GJ Dong, QL Zhan, CY Li, A Fujiyama, A Toyoda, TT Lu, Q Feng, Q Qian, JY Li, B Han (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497–501

Huang XH, XH Wei, T Sang, Q Zhao, Q Feng, Y Zhao, CY Li, CR Zhu, TT Lu, ZW Zhang (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–967

Ikehashi H (2009). Why are There *indica* type and *japonica* type in rice, History of the studies and a view for origin of two types. *Rice Sci* 1:1–13

Jiang X, WH Long, Q Peng, XC Zhang, XW Liu, DS Zhang, YL Gong, YS Zhang, SS Zhu (2018). Mapping of QTLs for heading date of rice based on SLAF_seq recombinant inbred lines population. *Mol Plant Breed* 16: 8068–8072

Jyothi B, B Divya, LV Subba Rao, P Laxmi Bhavani, P Revathi, P Raghuveer Rao, B Rachana, G Padmavathi, J Aravind Kumar, C Gireesh, MS Anantha, R Abdul Fiyaz, C Suvarna Rani, ARG Ranganatha (2018). New plant type trait characterization and development of core set among Indica and tropical japonica genotypes of rice. *Plant Genet Resour* 16:504–512

Li F, J Xie, X Zhu X Wang Y Zhao, X Ma, Z Zhang, MAR Rashid, Z Zhang, L Zhi, S Zhang, J Li, Z Li, H Zhang (2018). Genetic basis underlying correlations among growth duration and yield traits revealed by GWAS in rice (*Oryza sativa* L.). *Front Plant Sci* 9; Article 650

Li H, B Handsaker, A Wysoker, T Fennell, J Ruan (2009a). The sequence alignment-map format and SAMtools. *Bioinformatics* 25:2078–2079

Li R, C Yu, Y Li, TW Lam, SM Yiu, K Kristiansen, J Wang (2009b). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967

Li Y, XF Zeng, YC Zhao, JR Li, DG Zhao (2016). Identification of a new rice low-tiller mutant and association analyses based on the SLAF-seq method. *Plant Mol Biol Rep* 35:72–82

Liu LQ, QL Luo, W Teng, B Li, HW Li, YW Li, ZS Li, Q Zheng (2018). Development of *Thinopyrum ponticum*-specific molecular markers and FISH probes based on SLAF-seq technology. *Planta* 247:1099–1108

Lu BR, X Cai, X Jin (2009). Efficient indica and japonica rice identification based on the InDel molecular method: Its implication in rice breeding and evolutionary research. *Prog Nat Sci* 10:1241–1252

Magwa RA, H Zhao, W Yao, W Xie, L Yang, Y Xing, X Bai (2016). Genome wide association analysis for awn length linked to the seed shattering gene *qSH1* in rice. *J Genet* 95:639–646

Matsushita S, T Kurakazu, DK Sobrizal, A Yoshimura (2003). Mapping of genes for awn in rice using *Oryza meridionalis* introgression lines. *Rice Genet Newslett* 20:17–18

McKenna A, M Hanna, E Banks, A Sivachenko, K Cibulskis, A Kernytsky, K Garimella, D Altshuler, S Gabriel, M Daly, MA Depristo (2014). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genomics Res* 20:1297–1303

Price AL, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, D Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–911

Saeko K, I Takeshi, L Shao Yang, E Kaworu, F Yoshimichi (2006). An SNP caused loss of seed shattering during rice domestication. *Science* 312:1392–1396

Saitou N, M Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425

Shao GN, SQ Tang, J Luo, GA Jiao, XJ Wei, A Tang, JL Wu, JY Zhuang, PS Hu (2010). Mapping of qGL7-2, a grain length QTL on chromosome 7 of rice. *J Genet Genomics* 37:523–531

Shaun P, N Benjamin, TB Kathe, T Lori, MAR Ferreira, D Bender, J Maller, P Sklar, PIW de Bakker, MJ Daly, PC Sham (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Amer J Hum Genet* 81:559–575

Shen YJ, H Jiang, JP Jin, ZB Zhang, B Xi, YY He, G Wang, C Wang, L Qian, X Li, QB Yu, HJ Liu, DH Chen, JH Gao, H Huang, TL Shi, ZN Yang (2004). Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol* 135:1198–1205

Sheng XB, XF Wang, YN Tan, ZZ Sun, D Yu, GL Yuan, DY Yuan, MJ Duan (2019). Construction of qSH1 mutants in rice (*Oryza sativa*) using CRISPR/Cas9 and characteristic analysis of *mutagenesis*. *J Agric Biotechnol* 27:212–221

Song ZP, WY Zhu, J Rong, X Xu, JK Chen, BR Lu (2006). Evidences of introgression from cultivated rice to *Oryza rufipogon* (Poaceae) populations based on SSR fingerprinting: Implications for wild rice differentiation and conservation. *Evol Ecol* 20:501–522

Sun X, D Liu, X Zhang, W Li, H Liu, W Hong, C Jiang, N Guan, C Ma, H Zeng, C Xu, J Song, L Huang, C Wang, J Shi, R Wang, X Zheng, C Lu, X Wang, H Zheng (2013). SLAF-seq: An efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. *PLoS One* 8; Article e58700

Tamura K (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and macimum parsimony methods. *Mol Biol Evol* 28:2731–2739

Tetsuo O, M Hiroaki, O Taichi, Y Takuya, T Noriko, E Kaworu, Y Masahiro, E Takeshi, I Takeshi (2015). The Birth of a Black Rice Gene and Its Local Spread by Introgression. *Plant Cell* 27:2401–2414

Togninalli M, U Seren, JA Freudenthal, JG Monroe, DZ Meng, M Nordborg, D Weigel, K Borgwardt, A Korte, DG Grimm (2020). AraPheno and the AraGWAS Catalog 2020: A major database update including RNA-Seq and knockout mutation data for *Arabidopsis thaliana*. *Nucl Acids Res* 48:1063–1068

Toriba T, HU Hirano (2014). *DROOPING LEAF* and *OsETTIN2* genes promote awn development in rice. *Plant J* 77:616–626

Wang Y, J Li (2011). Branching in rice. *Curr Opin Plant Biol* 14:94–99

Xia C, LL Chen, TZ Rong, R Li, Y Xiang, P Wang, CH Liu, XQ Dong, B Liu, D Zhao, RJ Wei, H Lan (2014). Identification of a new maize inflorescence meristem mutant and association analysis using SLAF-seq method. *Euphytica* 202:1–10

Yang QH, HX Jin, Yu, XM Yu, XJ Fu, HJ Zhi, FJ Yuan (2020). Rapid identification of soybean resistance genes to soybean mosaic virus by SLAF-seq bulked segregant analysis. *Plant Mol Biol Rep* 38:666–675

Yang XH, XZ Xia, Y Zeng, BX Nong, ZQ Zhang, YY Wu, FQ Xiong, YX Zhang, HF Liang, GF Deng, DT Li (2018). Identification of candidate genes for gelatinization temperature, gel consistency and pericarp color by GWAS in rice based on SLAF-sequencing[J]. *PLoS One* 13:e0196690

Yuan J, X Wang, Y Zhao, NU Khan, Z Zhao, Y Zhang, X Wen, F Tang, F Wang, Z Li Z Li (2020). Genetic basis and identification of candidate genes for salt tolerance in rice by GWAS. *Sci Rep* 10; Article 9958

Zhang GQ (2020), Prospects of utilization of inter-subspecific heterosis between Indica and japonica rice. *J Integr Agric* 19:1–10

Zhang LB, QH Zhu, ZQ Wu, RI Jeffrey, SG Brandon, S Ge, T Sang (2009). Selection on grain shattering genes and rates of rice domestication. *New Phytol* 184:708–720

Zhao Z, H Gu, X Sheng, H Yu, J Wang, L Huang, D Wang (2016). Genome-wide single-nucleotide polymorphisms discovery and high-density genetic map construction in cauliflower using specific-locus amplified fragment sequencing. *Front Plant Sci* 7; Article 334

Zong Y, Z Chen, JB Innes, C Chen, Z Wang, H Wang (2007). Fire and flood management of coastal swamp enabled first rice paddy cultivation in east China. *Nature* 449:459–462