Theor Appl Genet (2011) 122:905–913
DOI 10.1007/s00122-010-1497-0

ORIGINAL PAPER

Functional markers developed from multiple loci in *GS3* for fine marker-assisted selection of grain length in rice

Chongrong Wang · Sheng Chen · Sibin Yu

Received: 2 June 2010 / Accepted: 4 November 2010 / Published online: 25 November 2010

Abstract The gene *GS3* has major effect on grain size and plays an important role in rice breeding. The C to A mutation in the second exon of *GS3* was reported to be functionally associated with enhanced grain length in rice. In the present study, besides the C-A mutation at locus SF28, three novel polymorphic loci, SR17, RGS1, and RGS2, were discovered in the second intron, the last intron and the final exon of *GS3*, respectively. A number of alleles at these four polymorphic loci were observed in a total of 287 accessions including Chinese rice varieties (*Oryza sativa*), African cultivated rice (*O. glaberrima*) and AA-genome wild relatives. The haplotype analysis revealed that the simple sequence repeats $(AT)_n$ at RGS1 and $(TCC)_n$ at RGS2 had differentiated in the wild rice whilst the C-A mutation occurred in the cultivated rice recently during domestication. It also indicated that A allele at SF28 was highly associated with long rice grain whilst various motifs of $(AT)_n$ at RGS1 and $(TCC)_n$ at RGS2 were mainly associated with medium to short grain in Chinese rice. The C-A mutation at SF28 explained 33.4% of the grain length variation in the whole rice population tested in this study, whereas $(AT)_n$ at RGS1 and $(TCC)_n$ at RGS2 explained 26.4 and 26.2% of the variation, respectively. These results would be helpful for better understanding domestication of *GS3* and its manipulation for grain size in rice. The genic marker RGS1 based on the motifs $(AT)_n$ was further validated as a functional marker using two sets of backcross recombinant inbred lines. These results suggested that the functional markers developed from four different loci within GS3 could be used for fine marker-assisted selection of grain length in rice breeding.

Communicated by T. Sasaki.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1497-0) contains supplementary material, which is available to authorized users.

C. Wang · S. Yu (✉)
National Key Laboratory of Crop Genetic Improvement
and College of Plant Science and Technology,
Huazhong Agricultural University, Wuhan 430070, China
e-mail: ysb@mail.hzau.edu.cn

S. Chen
School of Plant Biology, and International Centre
for Plant Breeding Education and Research,
The University of Western Australia,
Crawley, WA 6009, Australia

Introduction

Grain size, as specified by grain length, width, and length-to-width ratio, is a major determinant of grain appearance quality and grain weight in rice. It is an important agronomic trait for artificial selection in rice breeding. Breeders tend to select plants with large seed size for high yield and appropriate grain size for milling yield and market preferences. However, it is difficult for breeders to improve grain size efficiently by phenotypes, since the traits are quantitatively inherited (McKenzie and Rutger 1983). Identification of genes conferring the grain size variation will provide valuable targets in breeding application.

Rice varieties show large number of variation in grain size (Juliano and Villareal 1993). Many quantitative trait loci (QTLs) for grain size have been detected, of which four genes, grain size on chromosome 3 (*GS3*), grain weight on chromosome 2 (*GW2*), grain incomplete filling on chromosome 1 (*GIF1*), and seed width on chromosome 5 (*qSW5/GW5*), have been isolated and characterized recently (Fan et al. 2006; Song et al. 2007; Wang et al. 2008; Shomura et al. 2008; Weng et al. 2008). Particularly for grain length, the gene *GS3* has been identified with five exons encoding a putative phosphatidylethanolamine-binding protein (PEBP)-like domain, a transmembrane region, a putative tumor necrosis factor receptor (TNFR)/nerve growth factor receptor (NGFR) family domain and a von Willebrand factor type C (VWFC) module (Fan et al. 2006). A cysteine codon (TGC) to a termination codon (TGA) mutation (hereafter referred as C-A mutation) in the second exon of *GS3*, eliminated part of the PEBP-like domain and all the other three conserved domains. Further studies revealed that this C-A mutation was functionally associated with enhanced rice grain length (Fan et al. 2009; Takano-kai et al. 2009).

With the rapid development of different types of DNA markers, marker-assisted selection (MAS) has been playing a prominent role in plant breeding. However, there are still many challenges such as existence of any particular alleles in a given breeding line and availability of user friendly DNA markers for MAS application in complex traits (Young 1999; Xu et al. 2005). Both random genomic marker and genic marker could be used in MAS. The random genomic markers, however, are limited in MAS application due to their relatively low accuracy in selection caused by the genetic recombination between the marker and the target gene. The genic or functional markers, derived from polymorphic loci within genes affecting phenotypic variation, would overcome the problem of the recombination, and thus are highly predictive of phenotype, and will facilitate efficient selection of favorable alleles in breeding programs (Andersen and Lübberstedt 2003). Identification of agronomically important genes and mining of the alleles in natural populations are primarily required to develop the genic or functional markers (Andersen and Lübberstedt 2003; Takeda and Matsuoka 2008). In the case of *GS3*, a cleaved amplified polymorphic sequence (CAPS) marker has been developed based on the C-A mutation in the gene. This CAPS marker is highly associated with grain length, thus could be used for selection of rice grain length in breeding (Fan et al. 2009). However, its efficiency in MAS is still limited as the PCR product needs to be digested by a restriction endonuclease and this procedure is relatively expensive and elaborative once applied to a large breeding population. It would be useful to develop some PCR-based functional markers for grain length improvement.

Knowledge of the allelic diversity in *GS3* and their effects would be helpful for genetic manipulation of grain size in rice. Takano-Kai et al. (2009) demonstrated that the C-A mutation in *GS3* gene played a critical role in seed size differences among the modern subpopulations of *O. sativa*. However, they also observed that a large portion of the grain size variation among *O. sativa* could not be totally explained by the C-A mutation in *GS3*. Whether other polymorphic loci in gene *GS3* related to grain size exist in rice germplasm is still unclear. The objectives of the present study were to, (1) identify allelic variations at different loci within the gene *GS3* in a wide collection of rice germplasm including Chinese rice varieties (*O. sativa*), African cultivated rice (*O. glaberrima*) and wild relatives; (2) investigate the effects of various alleles and haplotypes for grain length; and (3) develop a set of functional markers based on the allelic variations at different loci within *GS3* for MAS in rice breeding program.

Materials and methods

Plant materials

Two hundred and eighty-seven accessions, including the mini-core collection of 213 Chinese landraces, varieties and elite parents of hybrid rice, which represented about 70% genetic diversity of Chinese rice germplasm (*O. sativa*) (Wen et al. 2009), 46 African cultivated rice (*O. glaberrima*), and 28 AA-genome wild rice accessions (*O. rufipogon*, *O. nivara*, *O. barthii,* and *O. meridionalis*), were used for allelic diversity assay at different loci of *GS3* gene (Table S1).

A set of backcross recombinant inbred lines (BRIL) developed from a cross between ‘Zhenshan97B’ (ZS97B) and ‘93-11’, and another set of BRIL and 34 near-isogenic lines (NIL) derived from a cross between accession IRGC 96717 (ACC9) and ‘ZS97B’ were used for functional marker validation. The *indica* cultivar ‘93-11’ and the *glaberrima* ‘ACC9’, both have long grains whilst *indica* parental line ‘ZS97B’ with short grains. For the BRIL of ‘ZS97B’ and ‘93-11’, the F_1 was backcrossed with ‘93-11’ to get BC_1F_1. Then a total of 242 BC_1F_1 plants were consecutively self-pollinated to produce BC_1F_8 by a single seed descendant method. For the BRIL of ‘ZS97B’ and ‘ACC9’, the F_1s backcrossed three times repeatedly to ‘ZS97B’ until BC_3 population obtained. A total of 100 BC_3F_1 plants were consecutively self-pollinated to produce BC_3F_5 by a single seed descendant method. The BC_3F_1 was also genotyped with genic markers for selection of ‘ACC9’ allele in *GS3* and with other 40 markers evenly distributed on rice chromosomes for screening genetic background. The genotypes carrying the ‘ACC9’ allele of *GS3* but otherwise background similar to ‘ZS97B’ were selected to backcross further with ‘ZS97B’ to obtain BC_4F_1. The BC_4F_2 was generated from the selfed BC_4F_1 and validated with the genic markers to construct near-isogenic lines. All accessions and lines were transplanted each in a single row with 10 plants at the experimental farm of Huazhong Agricultural University, Wuhan in rice growing season of 2006. Field managements followed essentially the normal agricultural practices. Grain length was evaluated as described previously (Fan et al. 2006). Ten randomly chosen grains of fully filled rice from each accession were lined up length-wise along a vernier caliper to measure grain length. The measurement repeated three times for each material.

Primers design

The simple sequence repeats (SSR) motifs were scanned by SSRHunter Version 1.3 (Li and Wan 2005) based on the genomic sequence of *GS3* that obtained from GenBank (accession number DQ355996). Primer pairs flanking the SSR motifs were designed using Primer Version 5.0 (http://www.premierbiosoft.com/primerdesign/). The primer sequences were aligned to the sequenced Nipponbare genome using BLAST to confirm their right locations in genome (http://www.ncbi.nlm.nih.gov). The CAPS marker SF28 was developed based on the C-A polymorphism in the second codon of *GS3* as described in the previous report (Fan et al. 2009). The primers that amplified the SR17 region were designed from the two allelic sequences of rice varieties 'Chuan 7' and '93-11' (Fan et al. 2006). All primers used in the study are listed in Table 1.

DNA extraction and PCR

DNA was extracted from fresh young leaves using CTAB method (Murray and Thompson 1980). PCR was performed in a total volume of 20 μL consisting about 30 ng DNA, 10 mM Tris-HCl, 50 mM KCl, 0.1%Triton X-100, 1.8 mM $MgCl_2$, 0.1 mM dNTP, 0.2 μM primers, and 1 U *Taq* DNA polymerase. Using a Gene AmpPCR system 9700 thermocycler (Perkin Elmer Cetus), PCR reactions were denatured at 94°C for 4 min, followed by 34 cycles of 94°C for 40 s, 55°C for 40 s and 72°C for 40 s. The final extension was at 72°C for 10 min. The PCR amplified productions were analyzed by electrophoresis in 6% polyacrylamide denaturing gels, and detected by silver staining (Ji et al. 2007). A 50 bp ladder marker (Fermentas) was used to estimate PCR fragment size. For the SF28, an amount of 8 μL PCR products was digested with 1 U *PstI* (TaKaRa, Dalian, China) according to the manufacturer's specification. The digested products were separated on the same detection system as mentioned above. The PCR products of SR17 were detected by electrophoresis in 1.0% agarose gels and stained by ethidium bromide.

Table 1 Four genic markers developed based on the polymorphism at four loci of *GS3* gene

Marker	Primer name	Primer sequence (5′–3′)
SF28	SF28F	TGCCCATCTCCCTCGTTTAC
	SF28R	GAAACAGCAGGCTGGCTTAC
SR17	SR17F	TGCCCATCTCCCTCGTTTAC
	SR17R	TGTTCGTTGCTGGTGTTG
RGS1	RGS1F	TCCACCTGCAGATTTCTTCC
	RGS1R	GCTGGTCTTGCACATCTCTCT
RGS2	RGS2F	GTGCATGATGCTTTCACCAC
	RGS2R	AGCGACACGGACTCTTCGT

Population structure analysis

The genetic structure of the mini-core collection of 213 accessions was investigated with the model-based method implemented in STRUCTURE (Pritchard et al. 2000) followed the procedure described in Agrama et al. (2007). Twenty-four SSR markers, evenly distributed on the 12 chromosomes of rice genome were selected to genotype the mini-core collection accessions for the population structure analysis.

Haplotype analysis and association analysis

PowerMarker version 3.1 was used to calculate allele frequencies and to construct the rooted phylogentic tree based on the polymorphic loci of *GS3* using neighbor-joining method (Cavalli-Sforza and Edwards 1967; Liu and Muse 2005; http://statgen.ncsu.edu/powermarker/). The tree was viewed by TREEVIEW (Page 1996). Association analysis was performed between grain length and the allelic variations with a general linear mixed model considering the population structure by TASSEL software (Bradbury et al. 2007).

Results

Polymorphic loci and their allelic diversity of *GS3* in rice germplasm

Except the C-A mutation in the second exon of *GS3*, three novel polymorphic loci were discovered. An insertion of 338 bp fragment in the second intron of *GS3* was detected in '93-11' as sequence comparison to 'Chuan 7'. A $(AT)_n$ motif in the last intron and a $(TCC)_n$ motif in the final exon of *GS3* were found in the genomic sequence of *GS3*. Three genic markers were newly developed based on these polymorphic loci: SR17, one insertion/deletion in the second intron of *GS3*; RGS1, a SSR marker with $(AT)_n$ motif in the last intron; and RGS2, another SSR marker with (TCC)n motif in the last exon of *GS3* (Fig. 1).

Using these three genic markers, together with the previously developed CAPS marker SF28 (Table 1), a total of 287 rice accessions were assayed and a number of allelic

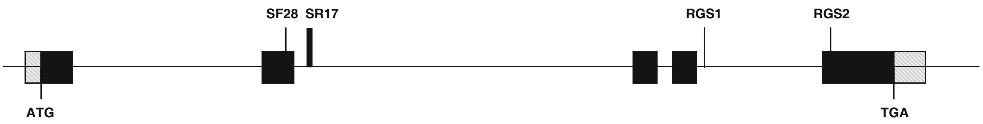

Fig. 1 The relative positions of the four polymorphic loci (SF28, SR17, RGS1, and RGS2) in *GS3*. The gene model is indicated with 5′ and 3′UTR (hatched boxes), exons (*black* boxes), introns (*lines* between the boxes), translation start codon (ATG) and translation stop codon (TGA)

variations were observed. For the SF28, all accessions generated PCR fragments of approximately 140 bp in size except of three wild rice (WR08, Y4, and Y5), which had the amplified fragments smaller than 140 bp. The PCR products of 239 accessions that might have the CTG**C**AG sequence at the *PstI* splicing site, could be digested by *PstI* to generate a 110 bp fragment, hereafter referred as C allele, while the amplification of the remaining 48 accessions that have the CTG**A**AG sequence did not be digested by *PstI*, scored as A allele (Table 2 and Fig. S1). Interestingly, the PCR products of the three wild relatives could also be digested by *PstI* (Table S1). The A allele with a frequency of 15.3% was observed in all the accessions, however, it was found neither in the wild rice, nor in the *O. glaberrima* surveyed (Tables 2, 3).

Three alleles were detected using the insertion/deletion marker SR17 (Table 2). A 1.1 kb segment was observed in high frequency of 81.9% in all accessions, which included 77 *indica*, 93 *japonica*, 20 wild rice and 45 *glaberrima*, while a 1.44 kb band detected in frequency of 17.4% in 50 accessions including 39 *indica*, 4 *japonica*, 1 *glaberrima* and 6 wild relatives (Table 2 and Fig. S1). Notably, a unique band of 800 bp was observed in two *O. meridionalis* (Y4 and Y5).

Table 2 Alleles and their frequencies at four polymorphic loci of GS3 in various populations sampled

Locus	Allele		Allelic frequency[a]				
	Name	Size (bp)	All accessions	Wild rice	*Glaberrima*	Sub 1	Sub 2
SF28	C	110	0.836	0.893	1.000	0.763	0.842
	6 bp del	104	0.007	0.071	–	–	–
	2 bp del	108	0.004	0.036	–	–	–
	A	140	0.153	–	–	0.237	0.159
SR17	−300bp	800	0.007	0.071	–	–	–
	1.1kb	1,100	0.819	0.714	0.978	0.672	1.000
	+338bp	1,438	0.174	0.214	0.022	0.328	–
RGS1	$(AT)_5$	180	0.379	0.321	0.022	0.748	0.010
	$(AT)_{12}$	194	0.453	0.571	–	0.251	0.988
	$(AT)_{13}$	196	0.160	0.036	0.978	–	–
	$(AT)_{85}$	340	0.007	0.071	–	–	–
RGS2	$(TCC)_3$	260	0.007	0.071	–	–	–
	$(TCC)_5$	266	0.537	0.321	1.000	0.748	0.012
	$(TCC)_6$	269	0.456	0.607	–	0.252	0.988

"–" Not observed

[a] Sub 1 and Sub 2 indicate the two subpopulations clustered for the 213 accessions by structure analysis

Table 3 Five main haplotypes across the four loci of *GS3* in rice

Haplotype	SF28 (+*PstI*)	SR17	RGS1	RGS2	*O. rufipogon*	*O. barthii*	*O. nivara*	*O. glaberrima*	*O. indica*	*O. japonica*	Total	Grain length[b]
H1	A	1.1kb	$(AT)_{12}$	$(TCC)_6$					28	16	44	9.09 ± 0.80 a
H2	C	1.1kb	$(AT)_{12}$	$(TCC)_6$	16				5	63	84	7.59 ± 0.77 b
H3	C	1.1kb	$(AT)_{13}$	$(TCC)_5$		1		45			46	8.50 ± 0.67 c
H4	C	1.1kb	$(AT)_5$	$(TCC)_5$	3				42	12	57	7.91 ± 0.65 d
H5	C	+338bp	$(AT)_5$	$(TCC)_5$	3		1	1	39	4	48	7.97 ± 0.60 d
Total					22	1	1	46	114	95	279[a]	

[a] Eight accessions in the haplotype with low frequencies are excluded

[b] The grain length of each haplotype is presented as mean ± SD. The different letters attached indicate that the means differences are significant at the level of 0.05 by Duncan's test

Four alleles were detected in the *GS3* using the marker RGS1 (Table 2). The size of the amplified products ranged from 180 to 340 bp representing 5, 12, 13, and 85 AT repeats, respectively (Fig. S1). The genotypes with $(AT)_5$ were observed mostly in *indica*, while the motif $(AT)_{12}$ was observed mostly in *japonica* (Table 3). Notably, the motif $(AT)_{13}$ was unique to the 45 *glaberrima* rice and its wild progenitor *O. barthii*. The two *O. meridionalis* produced 340 bp band as the genotype of $(AT)_{85}$.

Three alleles at the TCC motif amplified by the marker RGS2 were identified in the rice germplasm. A total of 128 accessions including 33 *indica*, 79 *japonica* and 16 wild rice had the $(TCC)_6$ allele with an about 269 bp band; the remaining 157 accessions including all 46 *glaberrima* samples carried the $(TCC)_5$ allele with a 266 bp band (Table 2 and Fig. S1). As revealed by the above markers, the two *O.meridionalis* contained a unique $(TCC)_3$ allele with about 260 bp band.

Haplotype variation

The allelic variations at the four loci of the *GS3* were used to construct gene haplotype. Nine haplotypes were revealed using the *O. meridionalis* (Y4) as out-group. Five haplotypes (H1–H5) were taken into account for polygenetic tree analysis (Table 3; Fig. 2), and the other four were excluded due to their low frequencies (<1%) across all accessions. As shown in Fig. 2, these five haplotypes were separated into two distinct clades. One class included two haplotypes, H1 and H2. The haplotype H1 contained 28 *indica* and 16 *japonica* that carry the mutant site A detected by SF28, and had long grain length (9.09 ± 0.80). Notably, no wild rice belonged to the haplotype H1. The haplotype H2, all with short grain length (7.59 ± 0.76), was found in about 73% wild rice, 66% *japonica* and 4% *indica* (Table 3). H1 had longer internal branch connecting the out-group than H2 (Fig. 2), suggesting that H1 might arise from H2. Another class included three haplotypes H3, H4, and H5. H3 was unique to the African species (*O. glaberrima* and *O. barthii*), which differed from H4 and H5 in the presence of $(AT)_{13}$ motif at the locus RGS1 (Table 3). The average grain length of H3 was 8.50 mm, which was significantly longer than that of H4, H5 and H2, while shorter than that of H1 (Table 3; Fig. 3). These results indicated that other loci in addition to C-A mutation within *GS3* contributed to the grain length variation in different cultivated rice.

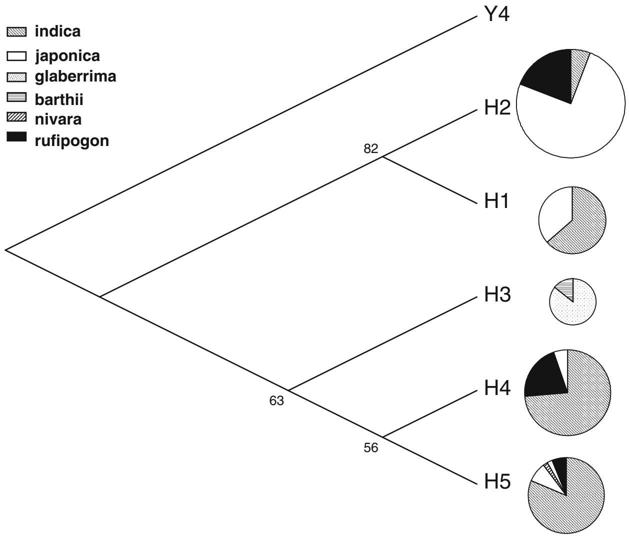

Fig. 2 Clustered haplotype tree constructed using neighbor-joining method for the 281 accessions that excluded 6 with the low frequency of haplotypes across the four polymorphic loci of *GS3*. The right pie paradigm indicates a proportion of different accessions types (*indica*, *japonica*, *glaberrima* and wild rice) in each haplotype (H1-H5). Y4 is the *O.meridionalis* accession as out-group. The labeled numbers by branches are bootstraps of 1,000 replicates

The phenotypic effect of each locus could be deduced from the paired haplotype comparisons as shown in Table 3 and Fig. 3. The comparison between H1 and H2 showing significant difference in grain length supported that the allele A at locus SF28 determined long grain length. H3 vs. H4, both have the same alleles at the three loci (SF28, SR17, and RGS2), indicated that the AT_{13}/AT_5 effect of RGS1 determined moderate or short grain. H4 versus H5 suggested that the allelic effect of SR17 was only marginal but not significant difference in grain length. Meanwhile, H2 versus H3/H4 showed the different phenotypic effects of the three allelic combinations at both loci RGS1 and RGS2, and determined moderate or short grain too.

Association between grain length and allelic variation

A predominant structure with two subpopulations (Sub 1 and Sub 2) in correspondence with *indica* and *japonica* subspecies was revealed in the panel of 213 accessions.

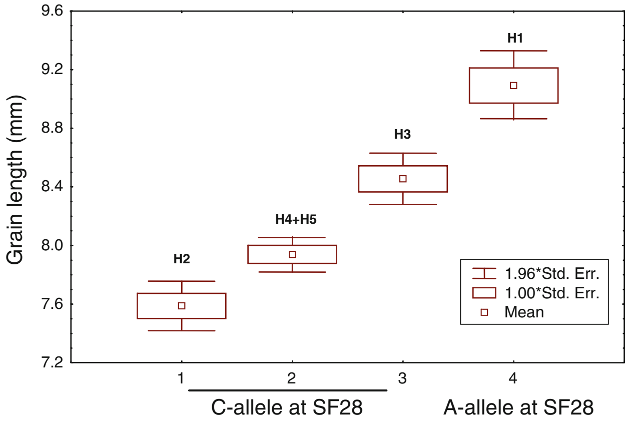

Fig. 3 The means of grain length with standard errors for the haplotypes showing significant differences among various haplotypes, particularly those carrying the C allele at locus SF28

Similar structure was also shown in the Chinese rice germplasm used by Wen et al. (2009). Thus, the Q value matrix with $K = 2$ was generated for further association analyses between grain length and allelic variations at the four polymorphic loci of *GS3*.

The C-A alleles, $(AT)_n$ and $(TCC)_n$ motif polymorphism at the loci (SF28, RGS1, and RGS2) were highly significantly associated with grain length among the mini-core collection of Chinese rice (Table 4). They could explain 33.4, 26.4, and 26.2% of the grain length variation, respectively, in the whole population, while they explained 52.1, 47.2, and 49.3% of the length variation in Sub 1. However, no associations were observed between grain length and SR17, RGS1 or RGS2 in Sub 2, because only one absolute predominance allele presented in this subpopulation (Table 2). As the five haplotypes with frequently more than 0.05 were entered the association analysis, high associations were also found between haplotypes and grain length (Table 4). It could explain 34.8, 18.9, and 52.5% of genetic variations in the whole population, Sub 1 and Sub 2, respectively.

Validation of haplotype effect involving RGS1 variation

The variation at RGS1 was primarily validated in the BRIL derived from the cross of '93-11' and 'ZS97B', which parental lines belong to haplotypes (H1 and H5) that carry the particular polymorphic alleles of $(AT)_{12}$ and $(AT)_5$ at RGS1, and the A and C alleles at SF28, respectively. As shown in Fig. 4a, the phenotype distribution of the BRIL was perfectly corresponded with their haplotypes of H1 and H5. Together with the A and C alleles at SF28, the genotypes of $(AT)_{12}$ and $(AT)_5$ at RGS1 were clearly co-segregated with the grain length variations in the BRIL. The 174 lines with H1 carrying $(AT)_{12}$ had an average of 9.2 ± 0.3 mm in grain length, which is significantly longer than that of the remaining 68 lines with H5 carrying $(AT)_5$ with an average of 8.2 ± 0.3 mm.

Table 4 Association analysis of grain length and four polymorphic loci in a panel of 213 rice accessions in China

Marker loci	Pop[a]	*F*	*P* value	R2 (%)
SF28	Whole	138.46	6.99E-25	33.4
	Sub 1	140.26	2.35E-22	52.1
	Sub 2	18.16	5.50E-05	18.5
SR17	Whole	11.72	7.41E-04	4.5
	Sub 1	9.61	2.40E-03	6.9
	Sub 2	–		No association
RGS1	Whole	96.19	6.19E-19	26.4
	Sub 1	115.53	1.23E-19	47.2
	Sub 2			No association
RGS2	Whole	94.98	9.40E-19	26.2
	Sub 1	125.63	8.87E-21	49.3
	Sub 2	–		No association
HAP[b]	Whole	29.04	2.99E-22	34.8
	Sub 1	6.05	9.27E-04	18.9
	Sub 2	27.59	9.71E-19	52.5

[a] Whole, Sub 1, and Sub 2 indicate the all 213 accessions and the two subpopulations clustered by structure analysis

[b] Haplotype

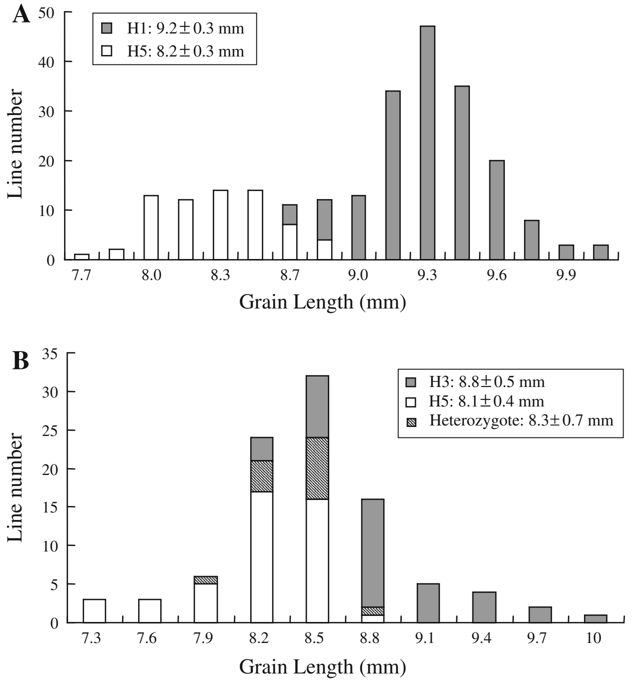

Fig. 4 Frequency distribution of grain length in two sets of backcross recombinant inbred lines, **a** derived from the cross of 'ZS97B' and '93-11' showing the effects of haplotypes H1 and H5, **b** derived from the cross of 'ZS97B' and 'ACC9', indicating that haplotypes H3 and H5 that carry the alleles $(AT)_{13}$ and $(AT)_5$ at the locus RGS1 consistent with short and long grains

The effect at RGS1 was further validated in the BRIL derived from the cross of 'ACC9' and 'ZS97B', where parental lines belong to haplotypes H3 and H5 that have the same C alleles at SF28, but contain the alleles $(AT)_{13}$ and $(AT)_5$, respectively at RGS1. The phenotype distribution of this BRIL was also corresponded with the haplotypes H3 and H5 (Fig. 4b). The 40 lines with H3 or $(AT)_{13}$ had an average of 8.8 ± 0.5 mm in grain length, which was significant by longer than that of the 45 lines carrying H5 or $(AT)_5$ with an average of 8.1 ± 0.4 mm. Since haplotypes H3 and H5 have the same C allele at SF28 (Table 3), it deduced that the grain length differences was caused by the alleles $(AT)_{13}$ and $(AT)_5$ at RGS1 in GS3.

The marker RGS1 also worked well in the development of the NIL derived from the variety 'ACC9' with long grain length. Significant difference was observed in grain length

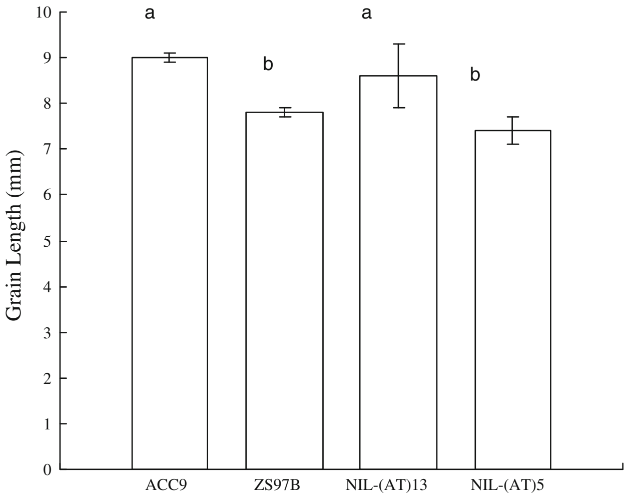

Fig. 5 Comparison of the average grain length among ‘ZS97B’, the near-isogenic lines (NIL) with or without the ACC9 haplotype carrying the allele $(AT)_{13}$ at RGS1. The standard deviation bars are indicated for each line and the *different letters* (**a**, **b**) attached represent that the means differences are significant at the level of 0.05

between the NILs which carried the haplotype of H3 or $(AT)_{13}$ and H5 or $(AT)_5$ detected by RGS1 (Fig. 5 and Fig. S2). The grain length of those 18 lines carrying $(AT)_{13}$ ranged from 8.1 to 10.0 mm with an average of 8.6 ± 0.7 mm; while the 16 lines containing $(AT)_5$ varied from 6.5 to 7.8 mm with an average of 7.4 ± 0.3 mm in grain length. These results indicated that the genic marker RGS1 could be used as a functional marker for selection of grain length.

Discussion

In the present study, three polymorphic loci, namely SR17, RGS1, and RGS2, were found on the second intron, the last intron and the final exon of *GS3*, respectively. Numerous insertion/deletion and nucleotide polymorphisms at these three loci were identified besides the C-A mutation at the locus SF28 within *GS3* in a wide collection of rice germplasm. Two features of the allelic diversity in *GS3* are noteworthy. The first is that the African species (*O. glaberrima* and *O. barthii*) hold a specific motif $(AT)_{13}$ at the locus RGS1 (Tables 2, 3), meanwhile the C-A polymorphism of *GS3* was detected in none of the African cultivated rice (*O. glaberrima)* tested in our study, nor in the other report (Takano-Kai et al. 2009), inferring that the African rice might have a domestication process different from the Asian cultivated rice (*O. sativa*). The second is that *O. meridionalis,* one of the wild relatives of rice, has unique alleles (or haplotype) across the four loci of *GS3*. This wild relative is found throughout northern Australia and shares the AA genome with *O. sativa* (Juliano et al. 2005), which makes it a good candidate for genetic improvement of the cultivated rice through introgressions of novel genes. In addition, there were four haplotypes in rare frequencies (<1%) observed in landraces and wild rice (Table S1), where one accession of *O. rufipogon* from Thailand conferring long length has two nucleotides deletion occurring at the locus SF28 that might disrupt PEBP-like domain. These allelic variations observed in the relatives of *O. sativa* would be valuable for domestication analysis and rice breeding. In the case of *GS3*, marker-assisted introduction of the distinctive alleles from *O. meridionalis*, *O. glaberrima* or *O. rufipogon* into the Asian cultivated rice is feasible using the developed markers in our study.

Another finding is that the allelic variations at the three loci including SF28, RGS1, and RGS2 in *GS3* were highly associated with grain length, and explained a large portion of the variations in the mini-core collection of Chinese rice germplasm (Table 3). The single nucleotide polymorphism (C-A) at the SF28 of *GS3* was confirmed to be highly associated with grain length in Chinese rice, consistent with the previous report that the sequence TGC occurred almost in the short-grain group and the TGA (coding stop) in the long-grain group (Fan et al. 2009). However, from the 287 accessions assayed in this study, 13 accessions with the TGC mutation did also show long grains (Table S1), suggesting that the C-A mutation could not totally explain the accessions with long or short grains in Chinese rice germplasm. Takano-Kai et al. (2009) has detected many nucleotide variations in the last intron and the final exon by sequencing the *GS3* gene in 54 diverse accessions of *O. sativa*, but they did not find any functional mutations other than the TGC/TGA mutant in the second exon. In the present study, two simple sequence repeats variations were detected in the last intron and the final exon of *GS3*, and these two loci could considerably affect grain length (Table 4), even if the TGC to the premature stop codon TGA did not happen in *GS3* (Fig. 3). There were also significant differences in grain length among the diverse haplotypes, H2, H3, H4, and H5. These results deduce that the sequence repeats variations might be individually or jointly functional in regulating of *GS3*. Many studies have reported that the $(CT)_n$ repeats or other motifs in genes such as starch biosynthesis genes *Waxy*, *SBEI* and *SS1* in rice affect their transcriptional activations (Tan and Zhang 2001; Bao et al. 2006). We developed a marker RGS1 based on the $(AT)_n$ motif in the last intron of *GS3*. Using RGS1 as a genic marker, we successfully obtained several NILs whose introgression segment contained the gene *GS3* from ‘ACC9’ with longer grains than that of non-*GS3* introgression lines (Fig. 5), and observed different transcripts of the gene with RGS1 variation in young panicles (data not shown), supporting the above deduction. We also developed another SSR marker RGS2 based on the $(TCC)_n$ motif in the last exon that encodes TNFR/NGFR domain of the *GS3*. The $(TCC)_n$ variations lead to the change of serine repeat number and thus might affect the binding ability of the protein. Taken together, other allelic variations besides the C-A allele in *GS3* should be causally responsible for the grain length variation in rice. Thus, RGS1 and RGS2, as complement markers with the CAPS marker SF28, have important implication for analyzing functional diversity in rice.

Origin of GS3 mutations

The grain size had played an important role in the evolution of rice and other cereal crops (Paterson et al. 1995; Kovach et al. 2007). Compared with their wild relatives, the grains of modern crops are usually longer and wider in shape, which means long grain is the product of artificial selection (Purugganan and Fuller 2009). It is widely recognized that domestication is not a single 'event', but rather a dynamic evolutionary process that occurs over time (Gepts 2004; Doebley et al. 2006). Takano-Kai et al. (2009) demonstrated that the A-allele for long grain was associated with strong positive artificial selection of *GS3* in tropical *japonica*, where this allele was observed in the highest allele frequency (61%) within *O. sativa*, while it is virtually absent from temperate *japonica*. In the present study, based on the allelic variations at the four loci of *GS3* in a panel of 287 accessions, most samples from Chinese rice germplasm, we provided further evidences supporting the short domestication history of the C-A mutation in *GS3* and the A allele for long grains evolved recently in cultivated rice. Moreover, we find that $(AT)_n$ motif at the last intron of *GS3* had diversification in the two cultivated rice species. Firstly, the C-A polymorphism was observed only in Asian cultivated rice, and the A-allele in *GS3* was found in none of the *O. rufipogon* and in low allele frequency in *O. sativa*. As comparison, the $(AT)_n$ and $(TCC)_n$ motifs at RGS1 and RGS2 had differentiation in wild rice (*O. rufipogon*) (Table 2), suggesting that the changes in these two motifs should be of earlier variation comparing to the C-A mutation. Secondly, two haplotype clusters was shown in Fig. 2. The first group included the haplotypes H1 and H2, which were mostly identical except for the C-A mutation that resulted in short/long grains. The haplotype H1 was far away from the outgroup than H2 that should be of original type existing in most of *O. rufipogon*. This pattern indicates that H1 might arise from H2. The A allele mutation, only existing in H1, was deduced to be derived from the haplotype H2 (Fig. 2). This is consistent with the long grain phenotype which occurred during rice domestication. In the other cluster, the haplotype H3 unique to the Africa rice (*O. glaberrima*), differed from H4 and H5 in the $(AT)_n$ motif (Table 2), suggesting that the African rice might have a domestication process different from the Asian cultivated rice (*O. sativa*). Thirdly, there were significant differences in grain length between *O. sativa* and *O. glaberrima* or among those *O. sativa* accessions carried the C allele in *GS3* (Fig. 3). In addition, there were four rare haplotypes (H6-H9), which might occur due to the recombination between RGS1 and RGS2 (Table S1). Thus, the variations of sequence repeat motifs at the loci RGS1 and RGS2 within *GS3* must be accumulated for the grain length difference before the effect of the C-A mutant observed. However, these alleles at the RGS1 and RGS2 have not been under strong artificial selection in rice, mainly due to their moderate effects on grain length relative to that of the C-A alleles within *GS3*. The C to A mutation can be considered a turning point in the grain length selection because it would have enhanced the grain length variation sharply in cultivated rice (Takano-Kai et al. 2009). Collectively, variations of the $(AT)_n$ and $(TCC)_n$ motifs at RGS1 and RGS2 should occurred earlier compared with the C-A mutation. These results would be helpful for better understanding of the domestication of *GS3* and allelic effects for grain size in rice.

Genic marker for grain length improvement

It is difficult for breeders to efficiently improve the quality traits such as grain length using conventional selection method, since these traits are quantitatively inherited and also largely affected by genotype–environment interaction. Moreover, the consumer-preferred long grain is controlled by the recessive allele of the major gene *GS3* in rice. Its heterozygote and dominant homozygote always express the short grain phenotype. It is impossible to select the genotype with long grains in the generation of hybrid based on phenotypic selection. Thus, it would be particularly helpful to develop genic markers of *GS3* for direct genotype selection in early generations of hybrid breeding program. A CAPS marker was developed previously based on the C-A polymorphism in *GS3*, and it was highly associated with grain length (Fan et al. 2009). This relationship is also confirmed by the present study (Table 4). The CAPS marker could be a functional marker for improvement of rice grain length. However, the polymorphism detection of this marker is subject to restriction digestion, and thus it is not as efficient as those PCR-based markers (such as SSR) in marker-assisted breeding practice. In the current study, two SSR markers, RGS1 and RGS2, are developed, the polymorphism of which are also highly associated with grain length in the mini-core collection of Chinese rice (Table 4). We can deduce the various effects of each locus in *GS3* from Table 3. For example, locus SF28 determines long/short grain length which is vital for long grain breeding; and the various allelic combinations at loci RGS1 and

RGS2 confer moderate/short grain and thus are important in rice breeding for medium or short grains to meet various demands of rice consumers; locus SR17 has only marginal effect to grain length, and probably not so important as SF28, RGS1 and RGS2 in rice breeding. These results are consistent with the genetic analysis in recombination inbred lines (Fig. 4) and the association analysis in germplasm (Table 4). However, further study, i.e., transgenic test, is still needed to validate the function of different alleles at each novel locus we reported here. The functional characterization of each single novel locus and the haplotype with or without C/A mutation in *GS3* by transgenic approach will be important and desirable for MAS of the gene. Anyway, these genic markers could be used instead of the marker SF28 for a large scale breeding application. For instance, they could be used for breeding application in genotyping the segregation populations derived from *O. glaberrima* or other rice relatives. Although medium length of rice seeds is a desirable processing trait for high yield of milled grains, it is not easy to select the medium grains based on phenotypic selection. Particularly, RGS1 is well predictive of medium to short grain length in rice (Table 3). Thus, it would be very useful in efficient selection of rice grains with appropriate length that is suitable for grain quality improvement.

Acknowledgments We thank Dr. F. Hu (Yunnan Agricultural Academy) for kindly providing most *O.glaberrima* accessions and the wild rice in Yunnan of China, Dr. X. Wei (China National Rice Research Institute) for most of the Chinese rice, and the International Rice Research Institutes for the wild rice relatives with IRGC accession numbers. This work was supported by grants from the National Program of High Technology Development of China, and the National Natural Science Foundation of China.

References

Agrama HA, Eizenga GC, Yan WG (2007) Association mapping of yield and its components in rice cultivars. Mol Breed 19:341–356

Andersen JR, Lübberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:554–559

Bao JS, Corke H, Sun M (2006) Microsatellite, single nucleotide polymorphisms and a sequence tagged site in starch-synthesizing genes in relation to starch physicochemical properties in nonwaxy rice (*Oryza sativa* L.). Theor Appl Genet 113:1185–1196

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

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309–1321

Fan CC, Xing YZ, Mao HL, Lu TT, Han B, Xu CG, Li XH, Zhang Q (2006) GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor Appl Genet 112:1164–1171

Fan CC, Yu SB, Wang CR, Xing YZ (2009) A causal C-A mutation in the second exon of *GS3* highly associated with rice grain length and validated as a functional marker. Theor Appl Genet 118:465–472

Gepts P (2004) Crop domestication as a long-term selection experiment. Plant Breed Rev 24:1–44

Ji YT, Qu CQ, Cao BY (2007) An optimal method of DNA silver staining in polyacrylamide gels. Electrophoresis 28:1173–1175

Juliano BO, Villareal CP (1993) Grain quality evaluation of world rice. International Rice Research Institute, Manila

Juliano AB, Naredo MEB, Lu BR, Jackson MT (2005) Genetic differentiation in *Oryza meridionalis* Ng based on molecular and crossability analyses. Genet Resour Crop Evol 52:435–445

Kovach MJ, Sweeney MT, McCouch SR (2007) New insights into the history of rice domestication. Trends Genet 23:578–587

Li Q, Wan JM (2005) SSRHunter: development of a local searching software for SSR sites. Yi Chuan 27:808–810 (in Chinese with an English abstract)

Liu K, Muse SV (2005) PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129

McKenzie KS, Rutger JN (1983) Genetic analysis of amylose content, alkali spreading score, and grain dimensions in rice. Crop Sci 23:306–313

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucl Acids Res 8:4321–4325

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Mol Biol 12:357–358

Paterson AH, Lin YR, Li ZK, Schertz KF, Doebley JF, Pinson SR, Liu SC, Stansel JW, Irvine JE (1995) Convergent domestication of cereal crops by independent mutations at corresponding geneticloci. Science 269:1714–1718

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature 457:843–848

Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M (2008) Deletion in a gene associated with grain size increased yields during rice domestication. Nat Genet 40:1023–1028

Song XJ, Huang W, Shi M, Zhu MZ, Lin HX (2007) A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nat Genet 39:623–630

Takano-Kai N, Jiang H, Kubo T, Sweeney M, Matsumoto T, Kanamori H, Padhukasahasram B, Bustamante C, Yoshimura A, Doi K, McCouch S (2009) Evolutionary history of *GS3*, a gene conferring grain length in rice. Genetics 182:1323–1334

Takeda S, Matsuoka M (2008) Genetic approaches to crop improvement: responding to environmental and population changes. Nat Rev 9:444–457

Tan YF, Zhang QF (2001) Correlation of simple sequence repeat (SSR) variants in the leader sequence of the waxy gene with amylose content of the grain in rice. Acta Bot Sin 43:146–150

Wang E, Wang J, Zhu X, Hao W, Wang L, Li Q, Zhang L, He W, Lu B, Lin H, Ma H, Zhang G, He Z (2008) Control of rice grain-filling and yield by a gene with a potential signature of domestication. Nat Genet 40(11):1273–1275

Wen W, Mei H, Feng F, Yu S, Huang ZC, Wu JH, Chen L, Xu XY, Luo LJ (2009) Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L.). Theor Appl Genet 119:459–470

Weng J, Gu S, Wan X, Gao H, Guo T, Su N, Lei C, Zhang X, Cheng Z, Guo X, Wang J, Jiang L, Zhai H, Wan J (2008) Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight. Cell Res 18:1199–1209

Xu Y, McCouch SR, Zhang Q (2005) How can we use genomics to improve cereals with rice as a reference genome? Plant Mol Biol 59:7–26

Young ND (1999) A cautiously optimistic vision for marker-assisted breeding. Mol Breed 5:505–510