Evolutionary history of GS3, a gene conferring grain length in rice

Noriko Takano-Kai*,1, Hui Jiang†1,2, Takahiko Kubo*3, Megan Sweeney†,‡, Takashi Matsumoto§, Hiroyuki Kanamori**, Badri Padhukasahasram††, Carlos Bustamante**, Atsushi Yoshimura*, Kazuyuki Doi*,4,6 and Susan McCouch†,5

*Faculty of Agriculture, Kyushu University, Fukuoka 812-8581, Japan
†Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853, USA
‡Department of Plant Sciences, University of Arizona, Tucson, AZ, USA
§Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan
**Institute of Society for Techno-Innovation of Agriculture, Forestry, and Fisheries, Tsukuba, 305-0854, Japan
††Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA
1 These authors contributed equally to this work.
2Present address: Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, Sichuan 610041, P. R. China
3Present address: Plant Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
4Present address: Graduate School of Bioagricultural Sciences, Nagoya University, Chikusa, Nagoya 464-8601, Japan
Running Head: History of GS3 in rice

Key Words: Rice (Oryza sativa L.), grain length, haplotype, LD decay, diversity, selection

5Co-Corresponding Author1: Susan McCouch, Cornell University, 162 Emerson Hall, Ithaca, NY USA 14853-1901; email: SRM4@cornell.edu; Phone: +1 (607) 255-0420; Fax: +1 (607) 255-6683

6Co-Corresponding Author2: Kazuyuki Doi, Nagoya University, Graduate School of Bioagricultural Sciences, Chikusa, Nagoya 464-8601. Email: kdoi@agr.nagoya-u.ac.jp

Article Summary: Domesticated rice exhibits dramatic phenotypic diversity for grain size and shape, unlike maize and wheat where artificial selection is associated with an almost uniform increase in seed or grain size. In this paper, we clone and characterize the gene, grain size 3 (GS3), a dominant regulator of grain size, and investigate its expression patterns. Using haplotype analysis, we demonstrate that the derived allele responsible for long grains originated in the Japonica varietal group and was transferred into Indica, despite the popular association of Indica varieties with the long-grained phenotype. These data provide the basis for genetic manipulation of grain size in rice and indicate that GS3 played a critical role in defining the seed morphologies that are characteristic of the modern subpopulations of O. sativa.

Sequence data from this article have been deposited with the DDBJ/GenBank Data Libraries under accession nos AB488612-AB488665 (under application).
ABSTRACT

Unlike maize and wheat, where artificial selection is associated with an almost uniform increase in seed or grain size, domesticated rice exhibits dramatic phenotypic diversity for grain size and shape. Here we clone and characterize GS3, an evolutionarily important gene controlling grain size in rice. We show that GS3 is highly expressed in young panicles in both short and long-grained varieties but is not expressed in leaves or panicles after flowering, and we use genetic transformation to demonstrate that the dominant allele for short grain complements the long grain phenotype. An association study revealed that a C to A mutation in the second exon of GS3 (A-allele) was associated with enhanced grain length in O. sativa but was absent from other Oryza species. LD was elevated and there was a marked reduction in nucleotide diversity (97%) across the gene in accessions carrying the A-allele, suggesting positive selection for long grain. Haplotype analysis traced the origin of the long-grain allele to a Japonica-like ancestor and demonstrated introgression into the Indica gene pool. This study indicates a critical role for GS3 in defining the seed morphologies of modern subpopulations of O. sativa and enhances the potential for genetic manipulation of grain size in rice.

INTRODUCTION

Seed size and seed number are the major determinants of crop yield in both the cereals and the grain legumes. Seed size was also a target of artificial selection during domestication, where large seeds are generally favored due to ease of harvesting and enhanced seedling vigor (Harlan et al. 1972). In rice, traits related to grain size and appearance have a large impact on market value and play a pivotal role in the adoption of new varieties (Champagne et al. 1999; Juliano 2003). However, different grain quality traits are prized by different local cultures and
cuisines and, unlike other cereals such as wheat, barley or maize that are sold largely in processed forms, the physical properties of rice grains are immediately obvious to consumers (FITZGERALD et al. 2009). Thus, rice offers a unique opportunity to investigate the genetics and evolutionary history of seed size and shape.

Cultivated rice (O. sativa) was domesticated in Asia from the wild progenitor O. rufipogon Griff. and/or O. nivara Sharma (CHEN et al. 1993; DALLY and SECOND 1990; ISHII et al. 1988; NAKANO et al. 1992; OKA 1988). Classical studies of the subpopulation structure of O. sativa have identified two primary sub-species or varietal groups, namely Indica and Japonica (OKA 1988; SUN et al. 2002; WANG and TANKSLEY 1989). Studies that have dated the divergence between the Indica and Japonica groups indicate that it predates rice domestication by at least 100,000 years (MA and BENNETZEN 2004; VITTE et al. 2004; ZHU and GE 2005), suggesting that at least two, genetically distinct gene pools of O. rufipogon were cultivated and subsequently domesticated.

Isozyme and DNA studies revealed that there is additional genetic structure within these two groups, with three subpopulations comprising the Japonica varietal group (temperate japonica, tropical japonica and aromatic, written all in lowercase) and two subpopulations comprising the Indica group (indica and aus) and (CAICEDO et al. 2007; GARRIS et al. 2005; GLASZMANN 1987; SECOND 1985). While there is great diversity of seed size and shape both within and between the different subpopulations of O. sativa, each subpopulation is popularly associated with a characteristic seed shape and size. Temperate japonica varieties are known for their short, round grains, indica and aus for slender grains, and within the aromatic subpopulation, (hereafter referred to as Group V varieties, according to the isozyme group designation (GLASZMANN 1987) the group of basmati varieties is highly valued for their very
long, slender grains (Juliano and Villareal 1993). Identification of the genes that control the range of seed size variation in rice will offer opportunities to study the evolutionary history and phenotypic diversification of the five subpopulations within *O. sativa* and also provide valuable targets for genetic manipulation.

In rice, four genes contributing to seed or grain size have been identified and characterized. The first, *grain size 3 (GS3)*, was isolated from an *indica x indica* population and found to encode a novel protein with several conserved domains including a PEBP (phosphatidylethanolamine-binding protein)-like domain, a transmembrane region, a putative TNFR (tumor necrosis factor receptor)/NGFR (nerve growth factor receptor) family domain, and a von Willebrand factor type C (VWFC) domain (Fan et al. 2006). A second gene, *grain weight 2 (GW2)*, was found to encode an unknown RING-type protein with E3 ubiquitin ligase activity (Song et al. 2007). The third, *grain incomplete filling 1 (GIF1)*, encodes a cell-wall invertase required for carbon partitioning during early grain-filling (Wang et al. 2008). Finally, the recently characterized *seed width 5 (SW5)* has no apparent homologue in the database but was shown to interact with polyubiquitin in a yeast two-hybrid assay, thus it likely acts in the ubiquitin-proteasome pathway to regulate cell division during seed development (Shomura et al. 2008; Weng et al.).

Many genes controlling seed size have also been identified in *Arabidopsis* and tomato, providing a framework for assembling the genetic pathway that determines this trait in dicotyledenous plants (Chaudhury et al. 2001; Jofuku et al. 2005; Li et al. 2008; Ohto et al. 2005; Orsi and Tanksley 2009; Roxrud et al. 2007; Schruff et al. 2006; Sundaresan 2005; Xiao et al. 2008; Yoine et al. 2006; Zhou et al. 2009). Several of these genes show maternal
control by regulating endosperm and/or ovule development (GARCIA et al. 2003; JOFUKU et al. 2005; OHTO et al. 2005; XIAO et al. 2008).

Numerous studies have identified rice QTLs associated with grain weight and grain length (www.gramene.org; (Ni et al. 2009)). Ten of these studies identified a seed size QTL located in the pericentromeric region of rice chromosome 3 using both inter- and intra-specific crosses (BRONDANI et al. 2002; KUBO et al. 2001; LI et al. 2004; LI et al. 1997; MONCADA et al. 2001; REDONA and MACKILL 1998; THOMSON et al. 2003; XIAO et al. 1998; XING et al. 2002; YU et al. 1997). In interspecific crosses, the wild accessions always contributed the dominant allele for small seed size at this locus. Comparative mapping of QTLs controlling seed weight in rice, maize and sorghum further suggested that orthologous seed size genes at this locus might be associated with domestication in all three crops (PATERSON et al. 1995).

In the current study, we used positional cloning and transformation to demonstrate that the GS3 gene underlies both the gw3.1 QTL (LI et al. 2004; THOMSON et al. 2003) and the lk3 QTL (KUBO et al. 2001). In transformation experiments, we demonstrated for the first time that the dominant allele for small grain size complements the long grain phenotype and we characterized the spatial expression patterns of the gene at different developmental stages. We undertook an association analysis to examine the relationship between the alleles at GS3 and the observed variation for grain length/size in both wild and cultivated rice. Finally, we examined sequence haplotypes across the GS3 region to look for evidence of selection, and to identify the origin of the mutation leading to increased grain length in O. sativa.
MATERIALS AND METHODS

**Fine-Mapping of gw3.1 and lk3:** A total of 4,148 BC5F2 individuals derived from a cross between Jefferson x *O. rufipogon*, and 1,641 BC3F3 plants derived from a cross between Asominori x IR24 were screened for recombinants within target regions defined by the QTLs, *gw3.1* and *lk3*, respectively. Plants from the Jefferson x *O. rufipogon* population were grown in 2-inch deep pots in the Guterman greenhouse of Cornell University, USA. Plants from the Asominori x IR24 population were grown at University Farm, Kyushu University, Japan. Once informative recombinants were identified, seedlings of recombinants and parental controls were transplanted and allowed to set seeds. The seed length phenotype was evaluated using seeds harvested from primary panicles. Seeds from the *gw3.1* population were measured using a digital caliper as described (LI et al. 2004) and seed lengths from the *lk3* population were determined by visual observation. Progeny testing was conducted as necessary.

Additional DNA markers were designed and used to detect recombination break points. Simple Sequence Repeat (SSR) markers were designed using the SSRIT tool ((TEMNYKH et al. 2000); [http://www.gramene.org/db/markers/ssrtol](http://www.gramene.org/db/markers/ssrtool)) and the publicly available Nipponbare genome sequence ([http://rice.plantbiology.msu.edu/](http://rice.plantbiology.msu.edu/)). SSR and indel markers were amplified using standard PCR protocols. Cleaved Amplified Polymorphic Sequence (CAPS) markers were designed based on Nipponbare sequences, PCR products were sequenced, and appropriate restriction enzyme were chosen; CAPS products were run on 2% agarose gels. Sequences of all primers are available in supplemental Table 1.
Complementation test and expression analysis of GS3: The dominant C-allele at GS3 was sub-cloned as a 7kb XmnI fragment of a Nipponbare bacterial artificial chromosome (BAC, OSJNHa0002D18) into the pPZP2H-lac binary vector (Fuse et al. 2001) (Figure 1). This construct was introduced into the chromosome segment substitution line AIS22 by Agrobacterium-mediated transformation (Toki 1997). AIS22 was constructed via backcrossing using Asominori (a short-seeded temperate japonica cultivar that is easy to transform) as the recurrent parent and IR24 as the donor of the long seed allele (Kubo et al. 2002).

Total RNA was extracted from samples using the Aurum total RNA mini purification kit (Bio-Rad, CA, USA). GS3 cDNA templates were generated from total RNA samples using the Revertra-Ace kit (TOYOBO, Osaka, Japan). The RT-PCR (reverse-transcribed PCR) products of the GS3 gene were sequenced. The sequences of the 5' and 3' ends of the cDNA were determined using the SMART RACE cDNA Amplification kit (Clontech, CA, USA).

To detect the differential expression of GS3 in wild type and mutant alleles, total RNA was extracted from 3 cm-long young panicles of cv. Asominori, AIS22 and cv. Nipponbare. To detect the time and spatial expression of GS3, total RNA was isolated from panicles at different developmental stages and leaves of Asominori for RT-PCR. GS3 was amplified using the primer pairs TGAGATCAAAAACTAGCTACTACCAGCTAGA and CATGGCAATGGCGGCGCGCAGCCGGCCCA. As controls, actin cDNA was amplified using the primer pairs TCCATCTTGCTCTCTTCAG and GTACCAGAGCAGGGATCT with fewer PCR cycles.

To confirm the time and spatial expression of GS3, a fragment that contained 1,030 bp of the promoter region of GS3 was amplified from genomic DNA of Nipponbare, cloned and fused with the beta-glucuronidase (GUS) gene and transferred into the temperate japonica variety,
Nipponbare. The panicles, leaves and leaf sheathes of the transgenic plants were stained to detect GUS activity as described by TAKEDA et al. (2003).

**Plant materials for the analysis of association, linkage disequilibrium (L.D), haplotype diversity and gene sequence:** Information about the germplasm used in this study is listed in supplemental Table 2. This study included 235 accessions of O. sativa from 30 countries (75 indica, 36 aus, 15 Group V, 34 temperate japonica, 64 tropical japonica and 11 admixed varieties), 79 accessions of O. glaberrima (SEMON et al. 2005), 10 O. barthii, 4 O. longistaminata, 12 O. glumaepatula, 12 O. meridionalis, 18 O. spontanea, and 266 O. rufipogon/O. nivara/O. spontanea. Accessions were surveyed to determine the frequency of the A-allele at GS3 and a subset was used for gene sequencing and haplotype analysis. The subpopulation identities of the O. sativa accessions were as determined previously (GARRIS et al. 2005); accessions new to this study were genotyped using 50 well-distributed SSRs (http://www.gramene.org/microsat/index.html) and analyzed using STRUCTURE as reported by GARRIS et al. (2005). Additional information is provided in supplemental Table 2.

**DNA extraction, PCR and sequencing:** DNA was extracted from leaf samples using a modified potassium acetate-SDS protocol (DELLAPORTA 1983). PCR was conducted using modified PCR protocols described previously (GARRIS et al. 2005) with the annealing temperature at 58°C. For sequencing, 3 µl of diluted PCR product was treated with 3ul 1:3 Exo-SAP (containing 3 units exonuclease I and 1.6 units shrimp alkaline phosphatase diluted with 1x PCR buffer) and incubated at 37°C for 45 min followed by 80°C for 20 min. Sequencing was performed using both forward and reverse primers to ensure accuracy on ABI Prism 3700/3100 DNA analyzers
Sequences were aligned using the Codon Code program (Codon Code, Dedham, MA). The ends of fragments were trimmed to remove low quality sequences. Heterozygous sites were identified by visual inspection of chromatograms for double peaks; the singletons and ambiguous sites were re-sequenced as necessary.

Supplemental Table 1 provides a list of all primers used for PCR and sequencing. Sequencing of the GS3 gene spanned the 5 exons and 4 introns, and also included 781 bp upstream and 16 bp downstream of the gene. The upstream border was set at -781bp because the genomic fragment used for transformation experiments extended 781 bp upstream and showed good complementation (Figure 1B); thus we believe our sequencing incorporates the promoter region.

**Association analysis:** For association analysis, 18-226 seeds from 157 accessions of *O. sativa* and 12-65 seeds from 162 accessions of *O. rufipogon/nivara/spontanea* (supplemental Table 2) were analyzed for seed length and seed width using the Winseedle scanner and software system (http://regentinstruments.com/). One CAPS marker, SHJ210 was designed to identify the functional mutation at GS3. PCR products amplified from the primer pair, SHJ210F (GCTTGATTTTCTGTGCTATTAGGAG) and SHJ210R (CTCAAAAAGCTTGACGATACGTATGT), were digested with the restriction enzyme, *Pst* I, and run on 2% agarose gels. Seed size and seed weight data analyses were conducted using JMP software (SAS Institute, Cary, NC).

**Sequence analysis of the GS3 Gene:** A total of 6.57 kb of DNA within the GS3 gene was sequenced in 54 diverse accessions of *O. sativa* (supplemental Table 2). Three pairs of PCR
primers were designed to amplify overlapping genomic regions within the gene and internal primers within each amplicon were designed to sequence the PCR products (supplemental Table 1). Sequences were assembled and aligned using the Sequencher program (Gene Codes, Ann Arbor, MI).

Nucleotide diversity ($\theta_\pi$) was calculated using the DnaSP program (ROZAS et al. 2003). Haplotypes were extracted using the same program after removing low-frequency alleles (<5%) and non-informative indels (i.e., poly A) to reduce the complexity.

**Haplotype and Genetic Diversity Analysis across the GS3 region:** Extended haplotypes (EH) spanning a 66.2 kb region flanking GS3 were used to distinguish the *indica* or *japonica* origin of the long-grain allele based on analysis of 172 accessions of *O. sativa* and 39 accessions of *O. rufipogon*. The haplotypes were constructed from SNPs and indels (frequencies >5%) identified in six 500-bp reads within a 66.2 kb region flanking GS3. Population genetic analyses were conducted using DnaSP 4.1 (ROZAS et al. 2003).

For Extended haplotype homozygosity (EHH) analysis, haplotypes were constructed from 269 SNPs and indels identified from fourteen 500-bp reads (supplemental Table 1) that spanned a 7 Mb region around GS3 (5 Mb downstream and 2 Mb upstream) in the same 172 accessions of *O. sativa* and 39 accessions of *O. rufipogon* as described above. The Fastphase program (SHEET and STEPHENS 2006) was used to fill in missing data to allow haplotype reconstruction across the target regions.
RESULTS

Fine mapping of gw3.1 and lk3: To determine whether the GS3 gene, which had previously been cloned from an indica x indica mapping population, was responsible for seed length and/or seed weight in crosses involving japonica cultivars, we fine-mapped two QTLs that had been previously mapped to the pericentromeric region of rice chromosome 3. The QTL gw3.1 (Thomson et al. 2003) was fine-mapped to a 22 kb region using a population derived from a cross between a long seeded tropical japonica (cv Jefferson) and an accession of O. rufipogon, and the QTL lk3 (Kubo et al. 2001) was fine-mapped to a 12 kb region in a cross between a short seeded temperate japonica (cv Asominori) and a long seeded indica (cv IR24) (Figure 1). The GS3 gene was contained within both fine-mapped regions and was polymorphic in the second exon (C165A) (Fan et al. 2006) in both pairs of mapping parents. These results were consistent with the hypothesis that a single mutation in GS3 was responsible for the seed size/seed weight QTLs in the peri-centromeric region of chromosome 3 in both indica and japonica genetic backgrounds. As documented by Li et al. (2004), seed size (caryopsis with hull) was highly correlated with grain size (caryopsis without hull, or brown rice grain; $R^2=0.975$) and we therefore refer to seed or grain size interchangeably in this paper.

Complementation test: To confirm that the GS3 gene was causally responsible for variation in seed or grain size in rice, we transformed a line containing the recessive C165A-mutation conferring long grain (hereafter referred to as the A-allele), with a 7 kb XmnI fragment containing the dominant short grain allele (the C-allele) from cv Nipponbare (Figure 1A). The recipient line used in this work was the chromosome segment substitution line, AIS22 (Kubo et al. 2002). AIS22 was genetically identical to the short-grained cultivar, Asominori, a temperate
*japonica* cultivar that is easy to transform, except that it contained an introgression from the long-grained *indica* cultivar, IR24, across the *GS3* region on chromosome 3. Due to this introgression, the grains of AIS22 were significantly longer (12.6%) and thinner (4.3%) than those of Asominori (p<0.01) (Figure 1B). No significant difference for grain thickness was observed between Asominori and AIS22. When AIS22 was transformed with the dominant, functional C-allele at *GS3*, 16 of 32 T₀ plants showed a short-grain phenotype (data not shown). To further confirm co-segregation between grain length and the transgene, we obtained two independent T₁ families from two T₀ plants and observed segregation patterns. The grain length of T₁ individuals containing the transgene was significantly shorter than the T₁ plants that did not inherit the transgene (Figure 1B; Table 1). We thus concluded that the wild type allele for short grains complemented the recessive long-grained phenotype.

**mRNA Expression of *GS3***: To determine in which tissues and developmental stages the gene was expressed, we examined the mRNA expression of *GS3* using RT-PCR. As seen in Figure 2A, *GS3* mRNA was expressed strongly in 3 cm-long panicles in both short-grained varieties (Asominori and Nipponbare) and in the long-grained chromosome substitution line (AIS22). As shown in Figure 2B, expression levels remained high as panicles developed from 3 – 7 cm in length, and then decreased as they elongated past 10 cm. At flowering, mRNA expression levels were below detection in the panicles and *GS3* mRNA was not detected in leaf tissue. This suggested that the *GS3* gene regulates grain size in rice during the early phases of panicle development while spikelets are elongating. Further experiments using a *GS3* promotor::GUS fusion confirmed this pattern of expression (Figure 2C, D). Because the genomic DNA fragment used for the complementation test contained only 781 bp upstream from the start codon, we used
a 1030 bp promoter for the GS3 promoter::GUS fusion. Another construct containing a 1984 bp promoter gave the same results. GUS expression was observed in panicles up until approximately 5 days before heading (Figure 2C), but the signal was not detected in either flowering panicles or leaves (Figure 2D).

**GS3 allele frequencies in diverse germplasm accessions:** The frequency of the A-allele conferring long grain was evaluated in 322 wild accessions, 235 accessions of *O. sativa* and 79 accessions of *O. glaberrima*. The A-allele was observed in 34% of *O. sativa* and 4% of *O. rufipogon/O. nivara/O. spontanea*, while it was observed in none of the other accessions (Table 2A, supplemental Table 2).

When GS3 allele frequencies were compared among the five subpopulations of *O. sativa*, highly significant (p<0.001) differences were observed (Table 2B). The A-allele was observed at relatively high frequency in *tropical japonica* (61%) and *Group V varieties*, (47%), but at low frequency in *temperate japonica* (6%). Within the *Indica* subspecies or varietal group (capitalized when referring to varietal group), the A-allele was present at moderately high frequency in *indica* (37%) (lower case when referring to subpopulation) but it was almost entirely absent from *aus* (3%).

**Phenotypic variation for seed morphology in wild and cultivated rice:** Using a subset of the wild (n=162) and cultivated materials (n=157) described above, no significant differences were observed in average seed length (p=0.1465) between the wild and cultivated groups, though seeds of *O. sativa* were significantly wider and heavier than seeds of *O. rufipogon* (p<0.001) (Table 3). Despite similar average seed lengths, the variance around the mean for seed length was
significantly greater in *O. sativa* than in *O. rufipogon* (p<0.0001). The coefficients of variation were also significantly greater for seed width, seed length/width ratio and seed weight (Table 3).

**Association between GS3 genotype and seed morphology:** When the two genotypic classes (A- and C-allele) of *O. rufipogon* were compared, there was no significant difference in seed length, width, length/width ratio or seed weight (Table 4). In contrast, *O. sativa* accessions carrying the A-allele (n=51) had significantly longer and thinner seeds than accessions carrying the wild type C-allele (n=106) (p<0.001) (Table 4; supplemental Figure 1). This suggested that GS3 affects seed morphology through interactions with factors in the genetic background that differs between *O. sativa* and its wild progenitor.

When the seed lengths of C-allele accessions of *O. sativa* were compared, there were significant differences among the subpopulations (Table 5). This provides evidence that genetic factors in addition to GS3 contribute to the variation in seed length in the different subpopulations of *O. sativa*. Paradoxically, A-allele accessions of *O. sativa* were all significantly longer than C-allele accessions, but there were no longer significant differences among the five subpopulations. This suggests that the A-allele of GS3 masks the differences in seed length that were detectable in accessions carrying the wild type allele.

When the five subpopulations of *O. sativa* were considered individually, the association between GS3 alleles and seed length was significant in every case, and R² values indicated that the A-allele explained 57% of the phenotypic variation for seed length in *Group V*, 27% in *indica*, 22% in *temperate japonica*, 15% in *tropical japonica* and 13% in *aus* (Table 4). GS3 was significantly associated with seed width only in the *indica* sub-group (R² = 0.26), and it was associated with length/width ratio in *Group V* (R² = 0.58) and *indica* (R² = 0.38) (supplemental
The association between GS3 alleles and seed weight was not significant in any of the individual subpopulations.

**Sequence and haplotype variation at GS3:** Sequencing of the GS3 gene from 54 diverse accessions of *O. sativa* identified a total of 86 SNPs and 28 indels in the 6.57 kb of aligned sequenced DNA (supplemental Table 4). Of these changes, two SNPs and one indel were in exons. Other than the C165A SNP in exon 2 (described above), neither of the other two polymorphisms in GS3 were associated with a significant difference in seed size, suggesting that they were not causally responsible for the phenotype.

*Japonica origin of the C165A mutation:* Given the reported magnitude of the Indica-Japonica differentiation in rice ($Fst = 0.47$, [Caicedo et al. 2007; Garris et al. 2005; Kovach et al. 2007]) we were interested to determine whether GS3 haplotypes showed evidence of divergent Indica and Japonica ancestry and if so, whether we could use ancestral differences to determine the origin of the C165A mutation.

Using the 70 SNPs/indels identified with allele frequency >5% in the 6.57 kb of GS3 sequence, we constructed a total of 14 gene haplotypes from 54 *O. sativa* accessions (Figure 3A). Considering only wild type (C-allele) accessions (33 accessions), a total of 11 ancestral Gene Haplotypes (GH1-11) were observed. To determine whether these Gene Haplotypes could be assembled into distinct clusters, we evaluated them using STRUCTURE and found the best resolved clusters at K=2 ([Pritchard et al. 2000]). Eighty-eight percent of the accessions from one cluster were from the Japonica varietal group, while 86% of accessions from the other cluster were from the Indica varietal group, defining the ancestral Japonica and Indica haplotype.
groups (Jap_GH and Ind_GH). Jap_GH contained two haplotypes (GH1, GH2); the haplotypes differed by only one SNP in intron 2 (indicated in yellow in Figure 3A). Ind_GH contained seven haplotypes (GH5-GH11). In addition, two accessions (haplotype GH3 and GH4) were classified as admixed because they shared ancestry with both Jap_GH and Ind_GH. The admixed accessions were recombinant haplotypes (Figure 3A).

Three Gene Haplotypes carrying the A-allele were identified (GH12- GH14) (Figure 3). GH12 clustered with Jap_GH, while GH13 and GH14 (recombinants) clustered with the admixed group. There were no A-allele haplotypes that clustered with Ind_GH. GH12, found in 90% of A-allele accessions, was identical to GH1 (C-allele) across the entire sequence of the GS3 gene, except for the functional C165A mutation (Figure 3A). Thus, we conclude that GH1, a Japonica haplotype, was the immediate ancestor of the C165A mutation. Our results demonstrated that all three A-allele haplotypes GH12, GH13 and GH14 are common by descent across the critical region of GS3 containing the functional SNP.

To test the hypothesis that a single, Japonica-derived mutation in GS3 was responsible for long grain in both the Indica and Japonica varietal groups, we examined a larger number of accessions (56 A-allele and 116 C-allele accessions of O. sativa) across a broad genomic region flanking GS3. A total of 30 extended haplotypes (EH1-30) were observed among the wild-type (C-allele) accessions, while only four extended haplotypes (EH31-34) were observed among the A-allele accessions. The extended C-allele haplotypes could be assembled into three distinct ancestral groups, corresponding to the composite Japonica varietal group (Jap_EH), and the two divergent subpopulations that comprise the Indica varietal group, indica (Ind_EH), and aus (Aus_EH) (Figure 3B).
Almost all of the A-allele accessions (91%) were found to carry the EH31 haplotype, which clustered with Jap_EH. This was entirely consistent with the situation described above for the Gene Haplotype that carried the A-allele, GH12. A single A-allele accession carried haplotype EH32, which differed from EH31 at a single SNP, and this accession also clustered with Jap_EH. Haplotypes EH33 and EH34 were represented by one and three accessions, respectively, and these four accessions were found to be recombinant types that clustered with Ind_GH. Despite clustering with Ind_GH, both EH33 and EH34 carry a region of Japonica-like DNA near the GS3 gene (Figure 3B). These results demonstrate that all accessions with the A-allele carry a genomic region flanking the GS3 gene that is closely related to ancestral Japonica. We can therefore conclude that this derived mutation in the GS3 gene conferring long grain arose only once in the Japonica gene pool or in a Japonica-like ancestor, and was disseminated through introgression into the Indica gene pool during the process of rice domestication.

**Origin of the C165A mutation within the Japonica varietal group:** To determine the origin of the A-allele within the Japonica varietal group, we examined polymorphism data in regions flanking GS3 from wild-type (C-allele) Jap_EH accessions looking for alleles that could differentiate between the temperate japonica, tropical japonica and Group V subpopulations.

An informative SNP that distinguished all eight ancestral (C-allele) Group V accessions from the other two Japonica subpopulations was identified 11 kb upstream of GS3 (Figure 3B). When this polymorphism was assayed in the 211 diverse accessions, it was not found in any other subpopulations of O. sativa, nor in any of the wild accessions in our study. None of the varieties carrying the A-allele, even the Group V accessions, contained this polymorphism. Therefore, we concluded that the C165A mutation did not originate in the Group V subpopulation, but must
have been introgressed from a different *Japonica* ancestor. We were unable to determine whether the mutation originated in *temperate* or *tropical japonica* due to the lack of polymorphism in the *GS3* region that could distinguish these subpopulations.

**Evidence for selection at GS3:** *O. sativa* accessions carrying the C165A mutation had a nucleotide diversity ($\theta_\pi$) of 0.0002 across the 6.57 kb of *GS3* sequence compared to $\theta_\pi=0.00464$ in wild type C-allele accessions. This 95.7% reduction in diversity was consistent with positive selection on the A-allele at this locus. We next investigated the extent of LD at *GS3* associated with A- and C-allele-accessions in the different varietal groups. In *O. rufipogon*, we observed rapid, symmetrical decay of LD around *GS3* in wild-type C-allele accessions, indicating a lack of selection at *GS3* in the wild progenitor (Figure 4). A similar pattern was observed for C-allele accessions in the *aus*, *indica* and *tropical japonica* subpopulations of *O. sativa*, with slightly slower LD decay in *temperate japonica* and in *Group V* accessions (consistent with a more intense domestication bottleneck in the latter two groups). In contrast, EHH extended over a larger region in all A-allele accessions, indicating an extended region of LD around *GS3* in these accessions (Figure 4). The patterns of LD and the marked reduction in $\theta_\pi$ observed in all A-allele accessions are indicative of strong positive selection for the derived allele conferring long grain at *GS3*. It is noteworthy that the pattern of EHH for A-allele accessions in *indica* and *tropical japonica* was almost identical, suggesting a very similar selection regime in these two groups.

**Gene flow between O. sativa and O. rufipogon at GS3:** Seven accessions of *O. rufipogon* carried the homozygous C165A mutation at *GS3*. To determine whether these wild accessions represented A-allele ancestors, or whether the A-allele had been transmitted as an introgression
from an *O. sativa* cultivar, we first compared the nucleotide diversity (θ,) of the seven A-allele and seven randomly-selected C-allele accessions of *O. rufipogon* based on concatenated sequences from the 66.2 kb region around the GS3 gene. The A-allele accessions had θ, = 0 while the C-allele accessions had θ, = 0.00883. This is similar to the reduction of genetic diversity (θ,) in A-allele vs. C-allele accessions of *O. sativa* in the GS3 gene region (0.0002 vs 0.00464, respectively). Further, when the A-allele haplotypes of *O. sativa* (EH31-32) were compared with A-allele haplotypes found in *O. rufipogon*, all seven wild accessions contained identical EH31 haplotypes. If the wild materials harbored an ancestral, pre-domestication version of the GS3 A-allele, we would expect to see greater diversity around GS3 in wild A-allele compared to cultivated A-allele accessions. However, because no sequence polymorphism was found in the regions flanking GS3 in A-allele accessions of *O. rufipogon*, and because all of the wild accessions harbored the same EH31 haplotype found in *O. sativa*, we concluded that the A-allele in these wild accessions were the result of recent introgression events from *O. sativa* to *O. rufipogon*.

**DISCUSSION**

**GS3 and the domestication process:** Unlike maize (DOEBLEY *et al.* 1994) and wheat (DUBCOVSKY and DVORAK 2007), rice domestication was not accompanied by a unidirectional increase in seed size, but rather by increased size variation, with artificial selection for seeds that were both longer and shorter, fatter and thinner than those of its wild progenitor, *O. rufipogon* (MORISHIMA *et al.* 1992). The fact that the A-allele had no effect on grain length in *O. rufipogon* suggests that it was not selected early in the domestication process. Other mutations were necessary before the phenotypic effect of the A-allele could be observed and we infer that these
other mutations accumulated in *O. sativa* prior to selection on the *GS3* locus. As such, the C165A allele can be considered a “diversification allele” because it would have enhanced the variation observed in *O. sativa* and contributed to the differentiation of the subpopulations within the cultivated gene pool. Evidence from this study suggests that once it attracted the attention of humans, it became a target of artificial selection and was introgressed from a *Japonica* ancestor(s) into the *Indica* gene pool. Interactions between the A-allele and diverse factors that distinguish the genetic backgrounds of the *Indica* and *Japonica* varietal groups would have generated novel grain morphologies and expanded the range of variation observed in *O. sativa*.

**Phenotypic impact of the A-allele within the *Japonica* and *Indica* varietal groups:** In this study, the A-allele at *GS3* contributes significantly to grain length in *tropical japonica, indica* and *Group V* varieties, and it confers unique grain morphologies in each of the subpopulations. In the *indica* and *Group V* backgrounds, the A-allele gives rise to long, slender grains that differ in appearance from the long, bold grains of *tropical japonica*. It will be of interest to determine whether the same genes confer slender grain in the *indica* and the *Group V* genetic backgrounds, or whether different alleles are responsible for grain width in these divergent subpopulations.

**Gene flow from cultivar to wild:** The C165A mutation is not found in the African cultivated species, *O. glaberrima*, nor in any of the wild species examined, except where it is associated with recent gene flow from *O. sativa*. Its presence in a few accessions of *O. rufipogon* is similar to the situation reported for the non-shattering allele, *sh4* (Li et al. 2006) and the *badh2.1* allele (M. Kovach, Cornell Univ., pers. comm.), except for the fact that the A-allele at *GS3* had no discernable phenotypic effect in the wild material.
**Origin within the Japonica and Indica varietal groups:** The history of the C165A mutation in GS3 suggests that it arose in a Japonica ancestor and moved into the Indica varietal group through introgressive hybridization. This pattern is reminiscent of the rc mutation for white pericarp and the Wxᵇ mutation for glutinous rice, that also arose in Japonica and became widely disseminated in the Indica gene pool over the course of rice domestication (KOVACH et al. 2007; SWEENEY et al. 2007; YAMANAKA et al. 2004).

In the case of GS3, the A-allele for long grain is associated with strong positive artificial selection in tropical japonica, where it attained the highest allele frequency (61%) of any varietal group within O. sativa, while it is virtually absent from temperate japonica. This suggests that it is likely to have arisen in the tropical japonica group where its presence may be used as a marker to help distinguish tropical japonica varieties from their close relatives in the genetically narrower temperate japonica group.

**Function of GS3 and the genetic pathway in seed size control:** Longer seed length in O. sativa is at least partially due to relaxed constraint on seed elongation mediated by the recessive C165A mutation in the GS3 gene. Results from both RT-PCR and GUS expression in transgenic plants showed that mRNA expression of GS3 begins during early panicle development (~3-5 cm), decreases when panicles are between 7-10 cm, and falls to below detection at flowering. In rice, panicles at the 3-5 cm stage are undergoing the early stages of inflorescence and ovule development, accompanied by the differentiation of glumes and floral organs. This period spans ovule primordial differentiation, integument primordial differentiation, division of integument primordium and meiosis of meristem mother cells (MMC) and early integument elongation (ITOH et al. 2005). Integument elongation begins in panicles that are 5-7 cm in length and
finishes when panicles reach 7-10 cm (ITOH et al. 2005), which corresponds precisely to the time when expression of GS3 starts to decline. This tissue- and stage-specific expression of GS3 suggests that it regulates seed size through control of ovule development.

In rice, the A-allele introduces a premature stop codon in the GS3 gene prior to the VWFC domain (FAN et al. 2006). Our RNA analysis showed that the expression pattern of the A-allele is similar to the wild allele, suggesting that the effect on phenotype is likely due to the truncation of the protein product itself, rather than to any difference in gene expression. We hypothesize that the wild type C-allele at GS3 functions as a dominant negative regulator of cell division and/or elongation in the integument.

The VWFC domain of the GS3 functional protein is reported to be important for protein-protein interaction and signaling (VAN VLIJMEN et al. 2004; ZHANG et al. 2007). Our genetic data demonstrates that the A-allele of GS3 masks the effects of other seed length genes in diverse accessions of O. sativa, and that it interacts differentially with seed width genes in the different subpopulations, supporting the hypothesis that it affects seed morphology via interaction with other genes. Identifying the interacting partners of GS3 will allow identification of how this gene affects seed size differently in diverse genetic backgrounds.

**Interaction between GW2 and GS3 does not explain subpopulation differences in seed size:**

Recently, a gene governing grain weight in rice, GW2, was cloned and found to encode a RING-type E3 ubiquitin ligase (SONG et al. 2007). A 1 bp deletion resulting in a premature stop codon in the GW2 gene was responsible for increasing seed size (SONG et al. 2007). The deletion increased the number of spikelet hull cells, which increased hull size and enlarged the endosperm cell size in mature rice grains. To investigate whether there might be an interaction between GS3
and GW2 that would help explain why GS3 affected grain size differently in different genetic backgrounds, we screened our germplasm panel to identify accessions that contained the 1bp deletion in GW2. Results of this survey demonstrated that none of the varieties of either O. sativa or O. rufipogon contained the reported functional mutation in GW2. Thus, the subpopulation differences in grain size observed in O. sativa are not the result of interaction between GS3 and GW2, but rather between GS3 and other, as yet unidentified, genes. The absence of the GW2 allele for large seed size in our panel suggests that it is either of very recent origin or it has been selected against by plant breeders and agriculturalists due to undesirable pleiotropic effects on grain quality. This is in direct contrast to the C165A mutation in GS3 that appears to have played a significant role in rice domestication and is found widely distributed throughout the rice growing world.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan to AY (Integrated Research Project for Plant, Insect and Animal using Genome Technology MP1125) and grants from the Plant Genome Program of the National Science Foundation of the USA (Award #0606461 and #0110004 to SMc). We thank Jiming Li for valuable advice and material used to fine map the gw3.1 QTL; Han Nguyen for assistance with genotyping and phenotyping early in the project; Lisa Polewsczak and Anna McClung at the USDA-ARS Rice Research Unit in Beaumont, Texas for generating the Winseedle data on seed size. We thank Yukihiro Ito for providing the GUS vector, Rod Wing and Jose Luis Goicoechea from the Arizona Genomics Institute for providing BAC clones and helping sequence through the GS3 region in O. rufipogon. We thank Lois Swales for help with
formatting and administrative support and Michael Kovach for critical review, discussion, and substantial editing of the manuscript.
LITERATURE CITED


OHTO, M., R. L. FISCHER, R. B. GOLDBERG, K. NAKAMURA and J. J. HARADA, 2005

Press, Tokyo.

associated with seed size evolution in tomato series. PLoS Genet. 5 e1000347.
doi:10.1371/journal.pgen.1000347.

Convergent domestication of cereal crops by independent mutations at

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure
using multilocus genotype data. Genetics 155: 945-959.

REDONA, E. D., and D. J. MACKILL, 1998 Quantitative trait locus analysis for rice panicle

ROXRUD, I., S. E. LID, J. C. FLETCHER, E. D. L. SCHMIDT and H.-G. OPSAHL-SORTEBERG,
2007 GAS4, One of the 14-Member Arabidopsis GASA Family of Small
Polypeptides, Regulates Flowering and Seed Development. Plant Cell Physiol. 48:
471-483.

ROZAS, J., J. C. SÁNCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA
polymorphism analyses by the coalescent and other methods. Bioinformatics 19:
2496-2497.


**TABLE 1.** Seed length and width of the transgenic plants in two independent T₁ lines. Means (mm ± SD of the plants with and without transgene were presented and compared by *t*-test. Parentheses indicate number of the plants. The average of the measurement of ten seeds was used as the value of each plant.

| Line | Seed length | | Seed width | |
|------|-------------|---|-------------|
|      | + transgene | no transgene | *P* | + transgene | no transgene | *P* |
| T₁ 1 | 7.2 ± 0.11 (7) | 8.2 ± 0.12 (4) | <0.01 | 3.2 ± 0.00 (7) | 2.9 ± 0.10 (4) | <0.01 |
| T₁ 2 | 7.2 ± 0.20 (6) | 8.3 ± 0.20 (4) | <0.01 | 3.1 ± 0.10 (6) | 2.9 ± 0.10 (4) | <0.01 |
TABLE 2. GS3 allele frequency in Oryza species (A) and subpopulations in O. sativa (B).

### A

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of accessions</th>
<th>A-allele</th>
<th>Freq. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O. sativa</td>
<td>235</td>
<td>81</td>
<td>34</td>
</tr>
<tr>
<td>O. nivara/O. rufipogon/O. spontanea</td>
<td>284</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>O. glaberrima</td>
<td>79</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O. barthii</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O. longistaminata</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O. glumaepatula</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>O. meridionalis</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### B

<table>
<thead>
<tr>
<th>Subpopulations in O. sativa</th>
<th>Varietal group</th>
<th>No. of accessions</th>
<th>A-allele</th>
<th>Freq. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O. sativa</td>
<td>Japonica</td>
<td>235</td>
<td>81</td>
<td>34</td>
</tr>
<tr>
<td>tropical japonica</td>
<td>Japonica</td>
<td>64</td>
<td>39</td>
<td>61</td>
</tr>
<tr>
<td>temperate japonica</td>
<td>Japonica</td>
<td>34</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>aromatic (Group V)</td>
<td>Japonica</td>
<td>15</td>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td>indica</td>
<td>Indica</td>
<td>75</td>
<td>28</td>
<td>37</td>
</tr>
<tr>
<td>aus</td>
<td>Indica</td>
<td>36</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>admixture</td>
<td></td>
<td>11</td>
<td>4</td>
<td>36</td>
</tr>
</tbody>
</table>
Table 3. Comparison of means and variances for seed morphology between *O. sativa* and *O. rufipogon* accessions (mean±SD)

<table>
<thead>
<tr>
<th></th>
<th><em>O. sativa</em></th>
<th><em>O. rufipogon</em></th>
<th>Means</th>
<th>Variances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=157</td>
<td>n=162</td>
<td>F ratio</td>
<td>P</td>
</tr>
<tr>
<td>Seed length (mm)</td>
<td>8.36 ±1.04</td>
<td>8.49 ±0.51</td>
<td>2.12</td>
<td>0.1465</td>
</tr>
<tr>
<td>Seed width (mm)</td>
<td>2.96 ±0.41</td>
<td>2.43 ±0.27</td>
<td>192.77</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SL/SW ratio</td>
<td>2.89 ±0.62</td>
<td>3.53 ±0.40</td>
<td>123.98</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Seed weight (g)c</td>
<td>23.21 ±4.82</td>
<td>15.75 ±3.16</td>
<td>269.07</td>
<td>&lt;0.0000</td>
</tr>
</tbody>
</table>

\(^a\)F ratio and probability based on one way ANOVA.

\(^b\)F ratio and probability based on O'Brien's test for equal variance.

\(^c\)Seed weight calculated as seed weight/1000 seeds.
**Table 4.** Comparison of means and variances for seed morphology between the C- and A-alleles at GS3 in *O. rufipogon* and *O. sativa*

<table>
<thead>
<tr>
<th></th>
<th>C-allele&lt;sup&gt;a&lt;/sup&gt;</th>
<th>A-allele&lt;sup&gt;b&lt;/sup&gt;</th>
<th>F ratio</th>
<th>P&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed length (mm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>O. rufipogon</em></td>
<td>8.49 ±0.50</td>
<td>8.49 ±0.65</td>
<td>0.41</td>
<td>0.52</td>
</tr>
<tr>
<td><em>O. sativa</em></td>
<td>7.94 ±0.90</td>
<td>9.21 ±0.75</td>
<td>9.86</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Seed width (mm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>O. rufipogon</em></td>
<td>2.43 ±0.27</td>
<td>2.36 ±0.27</td>
<td>0.41</td>
<td>0.52</td>
</tr>
<tr>
<td><em>O. sativa</em></td>
<td>3.03 ±0.39</td>
<td>2.82 ±0.40</td>
<td>9.86</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SL/SW ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>O. rufipogon</em></td>
<td>3.53 ±0.40</td>
<td>3.63 ±0.39</td>
<td>0.39</td>
<td>0.53</td>
</tr>
<tr>
<td><em>O. sativa</em></td>
<td>2.67 ±0.52</td>
<td>3.34 ±0.57</td>
<td>53.96</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Seed weight (g)&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>O. rufipogon</em></td>
<td>15.72 ±3.14</td>
<td>16.53 ±3.62</td>
<td>0.43</td>
<td>0.5</td>
</tr>
<tr>
<td><em>O. sativa</em></td>
<td>22.52 ±4.61</td>
<td>24.63 ±4.98</td>
<td>6.93</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<sup>a</sup>C-allele: *O. rufipogon* N= 155; *O. sativa* N = 106

<sup>b</sup>A-allele: *O. rufipogon* N= 7; *O. sativa* N = 51

<sup>c</sup>P-values calculated using the F-statistics

<sup>d</sup>Calculated as seed weight/1000 seeds
**TABLE 5.** Comparisons among *O. sativa* subpopulations for seed length between accessions carrying the A- versus C-allele at GS3 (mean±SD)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>C-allele</th>
<th>A-allele</th>
<th>R²</th>
<th>F ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seed length (mm)</td>
<td>N</td>
<td>Level&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Seed length (mm)</td>
<td>N</td>
</tr>
<tr>
<td>tropical japonica</td>
<td>8.47 ±0.85</td>
<td>14</td>
<td>A</td>
<td>9.18 ±0.83</td>
<td>27</td>
</tr>
<tr>
<td>temperate japonica</td>
<td>7.64 ±0.59</td>
<td>27</td>
<td>BC</td>
<td>9.10 ±2.21</td>
<td>2</td>
</tr>
<tr>
<td>aromatic (Group V)</td>
<td>7.21 ±1.33</td>
<td>8</td>
<td>C</td>
<td>9.55 ±0.54</td>
<td>6</td>
</tr>
<tr>
<td>indica</td>
<td>8.23 ±0.86</td>
<td>24</td>
<td>AB</td>
<td>9.11 ±0.46</td>
<td>15</td>
</tr>
<tr>
<td>aus</td>
<td>7.92 ±0.89</td>
<td>33</td>
<td>ABC</td>
<td>9.94 ±0.64</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup>Probability calculated using the F-statistics

<sup>b</sup>Calculated as seed weight/1000 seeds

<sup>c</sup>Multiple means comparisons based on Tukey-Kramer Honestly Significant Difference (HSD) test. Levels not connected by same letter are significantly different.
FIGURE LEGENDS

FIGURE 1. Fine mapping of grain size QTLs from two independent populations.
(A) Location of fine mapped regions on short arm of rice chromosome 3. Mb positions are based on TIGR pseudomolecule v.5; SSR and indel markers delineate fine-mapping of gw3.1 in Jefferson x O. rufipogon population and lk3 in Asominori x IR24 population; Xmn1 fragment contains GS3 gene used to transform AIS22; GS3 gene model shown 5’ to 3’ orientation.
(B) Seed phenotypes (photos) and mean length and width for parents of two mapping populations: Jefferson x O. rufipogon and IR24 x Asominori; AIS22= japonica line carrying A-allele used for transformation; T1 line= transformed with the C-allele. Bar=5mm.

FIGURE 2. Expression analysis of GS3 using RT-PCR and genetic transformation.
(A) RT-PCR of GS3 from Asominori, AIS22 and Nipponbare using young panicles (~3 cm) for RNA extraction. Numbers in parentheses indicate cycles of PCR. Actin cDNA was amplified as a control.
(B) RT-PCR of GS3 from Asominori using panicles approximately 3 cm, 5 cm, 7 cm, 10 cm, 13 cm and after heading, and a flag leaf for RNA extraction.
(C) and (D) GUS staining of transgenic plants transformed with the GS3 promotor:GUS fusion construct showing panicles before flowering (C) and a panicle at flowering with a flag leaf (D). Bars = 1cm.

FIGURE 3. Haplotype data across the GS3 gene region.
(A) The gene model for GS3 containing five exons and comprising approximately 6.2 kb is shown horizontally along the top; the position of the C165A mutation is shown in red; SNP
positions with $GS3$ are connected by lines to the haplotype table below. Rows in the table correspond to gene haplotypes (GH#) which are organized into groups based on STRUCTURE analysis; columns indicate the distribution of polymorphisms at each SNP position; blue indicates the common Group 1 SNP allele; yellow indicates a variant SNP allele; red highlights the presence of the A-allele for long grain at the C165A SNP. The number and subpopulation identity of varieties carrying each haplotype are indicated in columns to the right.

(B) Extended haplotype (EH#) table corresponding to a 66.2 kb region flanking the $GS3$ gene; position of the C165A mutation within the extended haplotype indicated in red; colors are the same as described in (A) above; stippled yellow squares indicate presence of a third allele; SNP names and primers used to detect them (supplemental Table 1) are shown along the $x$-axis at the bottom.

**FIGURE 4.** EHH across the $GS3$ genomic region for *O. rufipogon* and the five varietal groups of *O. sativa*. EHH values for individuals containing the C-allele (short grain) or A-allele (long grain) at $GS3$ indicated by thin or bold lines, respectively. Arrow indicates relative position of the C165A mutation; S# designations along the $x$-axis indicate primers (supplemental Table 1) used to detect polymorphisms; tick marks along the $x$-axis indicate correspondence between polymorphisms and amplicons.
SUPPLEMENTAL MATERIALS

**Supplemental Table 1.** List of the primers used in this study

**Supplemental Table 2.** Germplasm accession information and relationship to phenotype and haplotype

**Supplemental Table 3.** Associations between GS3 alleles and seed morphology (mean±SD). (A) seed width, (B), seed length / width ratio (C) and seed weight in *O. rufipogon* and *O. sativa* varietal groups

**Supplemental Table 4.** Numbers of SNPs and indels within GS3 based on sequencing of 54 accessions of *O. sativa*

**Supplemental Figure 1:** Distribution of seed size and seed weight in *O. rufipogon* and five varietal groups of *O. sativa*
A

Chromosome 3

16.60
16.65
16.70 Mb

centromere


IND120 gw3.1 HJ40

13-469 lk3 13-12726

XmnI fragment used for transformation

3' GS3

ATG C165A

5'

3' TGA

B

<table>
<thead>
<tr>
<th>GS3 allele</th>
<th>Seed length (mm±SD)</th>
<th>Seed width (mm±SD)</th>
<th>Seeds measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jefferson</td>
<td>A-allele 8.29±0.41</td>
<td>2.71±0.18</td>
<td>32</td>
</tr>
<tr>
<td>O. rufipogon</td>
<td>C-allele 8.04±0.34</td>
<td>2.98±0.13</td>
<td>35</td>
</tr>
<tr>
<td>IR24</td>
<td>A-allele 9.14±0.37</td>
<td>2.62±0.05</td>
<td>10</td>
</tr>
<tr>
<td>Asominori</td>
<td>C-allele 7.56±0.19</td>
<td>3.25±0.09</td>
<td>10</td>
</tr>
<tr>
<td>AIS22</td>
<td>A-allele 8.51±0.13</td>
<td>3.11±0.08</td>
<td>10</td>
</tr>
<tr>
<td>Transgenic (T1)</td>
<td>A-allele &amp; C-allele 7.19±0.11</td>
<td>3.16±0.05</td>
<td>70</td>
</tr>
</tbody>
</table>
Figure 2
Figure 3

A

B

Admixture/Recombinants

Aus E H

Ind EH

Ind GH

Admixture/EH

Jap_GH

Ind_Jap_EH

Group V

tropical japonica
temperate japonica
indica aromatic

Aus admixture
Takano-Kai / Jiang et al.

FIGURE 4

- O. rufipogon
  - C: n=37

- aus
  - A: n=1
  - C: n=15

- indica
  - A: n=14
  - C: n=35

- trop. japonica
  - A: n=28
  - C: n=23

- temp. japonica
  - A: n=2
  - C: n=29

- aromatic (Group V)
  - A: n=7
  - C: n=8