Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice

Gopal Misra1, Saurabh Badoni2, Roslen Anacleto1, Andreas Graner3, Nickolai Alexandrov2 & Nese Sreenivasulu1

In this study, we used 2.9 million single nucleotide polymorphisms (SNP) and 393,429 indels derived from whole genome sequences of 591 rice landraces to determine the genetic basis of cooked and raw grain length, width and shape using genome-wide association study (GWAS). We identified a unique fine-mapped genetic region GWi7.1 significantly associated with cooked and raw grain width. Additionally, GWi7.2 that harbors GL7/GW7 a cloned gene for grain dimension was found. Novel regions in chromosomes 10 and 11 were also found to be associated with cooked grain shape and raw grain width, respectively. The indel-based GWAS identified fine-mapped genetic regions GL3.1 and GWi5.1 that matched synteny breakpoints between indica and japonica. GL3.1 was positioned a few kilobases away from GS3, a cloned gene for cooked and raw grain lengths in indica. GWi5.1 found to be significantly associated with cooked and raw grain width. It anchors upstream of cloned gene GWS, which varied between indica and japonica accessions. GWi11.1 is present inside the 3′-UTR of a functional gene in indica that corresponds to a syntenic break in chromosome 11 of japonica. Our results identified novel allelic structural variants and haplotypes confirmed using single locus and multilocus SNP and indel-based GWAS.

Rice (Oryza sativa L.) being the most important food crop for more than half of the world’s population fulfills 45–70% of the daily caloric requirement of rice consumers in Asia (http://ricepedia.org/rice-around-the-world/asia). Rice accumulated substantial genetic variation during domestication leading to differences in seed morphology. This rice domestication process that occurred in various rice-eating cultures in Asian societies has led to create substantial genetic variability between indica and japonica subspecies1,2. Specific allelic combinations of agronomically important genes have been selected over others that led to huge phenotypic variation for traits including grain dimensions and shape at the subspecies level3-8. The key grain quality traits that factor into different consumer preferences are grain length, width, and grain shape (ratio between grain length and width). Variation in indica and japonica subspecies is also evident through the specific conserved allelic variants of major genes such as grain size 3 (GS3) and grain width 5 (GWS)9-11. Based on observed inferences of grain dimensional characteristics, indica rice must have undergone positive selection for long grains while japonica for relatively shorter and bold grains12. In the case of tropical japonica (also termed as javanica rice), they have larger grains than that of temperate japonica varieties13,14. The genetic basis of huge variations for grain dimensions and shape at the subspecies level arose from chromosomal rearrangements and duplications resulting in altering syntenic relationship between the subspecies9,15,16.

The evolutionary and domestication histories of rice made it best suited for GWAS. This technique has been extensively used in rice as an efficient strategy for the genetic analysis of complex traits that include grain dimensions3,5,6,8,12,16. High-resolution dissection of universal- and population-specific large effect alleles in GWAS aids in the understanding of locally adapted allele complexes in different subspecies5. In addition, targeted gene

1Grain Quality and Nutrition Center, Plant Breeding Division, International Rice Research Institute, DAPO Box 7777, Metro Manila, 1301, Philippines. 2Genetics and Biotechnology Division, International Rice Research Institute, DAPO Box 7777, Metro Manila, 1301, Philippines. 3Leibniz institute of Plant Genetics and Crop Plant Research (IPK), Correnstrasse 03, 06466, Gatersleben, Germany. Correspondence and requests for materials should be addressed to N.S. (email: n.sreenivasulu@irri.org)
association study (TGAS) has emerged as a complementary strategy to harvest other allelic variations within the candidate genes underlying genomic regions that were implicated in preceding GWAS analyses. Hence, combining GWAS with an integrated approach of targeted haplotyping on candidate gene from TGAS is considered an effective strategy for the genetic dissection of complex quantitative genomically important traits in rice.

Published GWAS on rice traits had always been on SNP-based genotyping data that practically missed out on the importance of capturing contributions of structural variations across sub-species in rice, which may be able to explain a fraction of missing heritability computed from mixed linear model SNP-based GWAS.

The grain dimensions and shape after cooking are among the major determinants of consumer preference, especially for rice-eating cultures that prefer non-bold varieties. Very limited studies have been conducted to identify the genetic causation of grain dimensions and shape in cooked rice. Hence, this study was focused on providing novel genetic information that could help explain further the variation in cooked and raw grain length, width, and shape. Complementing GWAS with TGAS using high-resolution SNP- and indel-based genotyping data on 591 diverse germplasm from the 3,000 Rice Genomes was seen as a robust and reliable analytical framework for the genetic dissection of these traits. Novel allelic variants identified from fine-mapped genetic regions on chromosomes 5 and 7 formed haplotypes that explained specific phenotypic ranges for cooked and raw grain width, while that in chromosome 11 specifically for raw grain width and chromosome 10 for cooked grain shape. Synteny between indica and japonica subspecies analyzed on the genomic regions implicated through GWAS and TGAS revealed rich structural variations that deepen the genetic understanding of grain shape related traits in these perversely cultivated rice subspecies. These haplotypes and structural variants will be useful targets for breeding programs to address the shape and dimensions of cooked rice.

Results

Phenotypic variation of diversity panel for cooked and raw grain length, width and shape traits. A wide range of values for grain length, width, and shape were observed both for raw and cooked grains in the diversity panel (Fig. S1). Unlike breeding lines, the frequency distributions of the grain dimensions and shape of indica landraces was not only enriched primarily for grain length but depicted variability for favored bold grains as well. In indica accessions, cooked grain length (GLc) ranged from 7.0 to 14.0 mm and raw grain length (GL) ranged between 4.31 to 7.56 mm with the average value of 5.96. Cooked grain width (GWi) was narrower in indica ranging from 2.5 to 4.25 mm, and raw grain width (GWi) was between 1.79 to 3.09 mm with the mean value of 2.43 mm (Fig. S1A,B). Among japonica genotypes, GLc ranged from 8.0–15.0 mm and GL from 4.1 to 7.33 mm with mean 5.73 mm. The phenotype variability in japonica for GWi, ranged from 3.0 to 4.6 mm and GWi from 1.87 to 3.31 mm with the mean 2.71 mm (Fig. S1C,D). For cooked grain length-to-width ratio (grain shape, GS), the means across indica and japonica accessions were 3.0 and 2.85 mm, respectively (Fig. S1). The phenotypic variations in indica germplasm panel were approximately normally distributed for all three traits. However, skewed phenotypic distributions were observed in GWi and GS in japonica subspecies so data were transformed prior to GWAS (Fig. S2A–D).

Identification of cooked and raw grain length, width and shape associated genetic variants through SNP-based genome wide association study. The high-quality re-sequencing data generated from 591 gene bank accessions composed of indica, temperate and tropical japonica subspecies from 72 countries representing global genetic diversity were used in this study. SNPs were called using the japonica (Nipponbare release 7) reference genome. Genetic structure and linkage disequilibrium estimation in rice germplasm panel was assessed. The mean SNP density was approximately one SNP at every 127 base pairs (or 8,053 SNPs/kb) across the rice genome. A total number of 2,260,030 SNPs were identified within indica, 1,562,078 from tropical and temperate japonica sub-groups and 2,933,037 from the population formed by merging indica and japonica genomic data (Supplementary note). All these quality-assured SNPs sets were used to calculate GWAS against cooked grain length, grain width, grain shape (GLc, GWi, and GS), and raw grain length, grain width, grain shape (GL, GWi and GS) within each subspecies and across the whole germplasm panel combining both subspecies. Both kinship and inferred population structure were used as covariables in a mixed linear model-based GWAS analysis. We used single-locus (SL)-GWAS approach using EMMAX for association analysis where Bonferroni corrected P-values with log10(P-value) > 5 were used as a threshold criterion to fetch moderate to highly significant loci. Furthermore, we adopted the multi-locus (ML)-GWAS strategy by following three independent methods. We detected the common highly significant hotspots validated using single and multi-locus methods on the characteristic genetic regions on chromosomes 3, 5, 7 and 11 (Table S1). Significant associations with GS were observed on chromosomes 3, 5, 7 and 10 that associated with either GLc or GWi (Figs 1–3). Chromosome 3 loci (GLc.1 and GLc.1.), respectively associated with GL and GWi, and GS (Figs 1–3, Figs S3 and S4). While, a number of significant association signals were detected in the genomic regions on chromosome 5 (GWS) and 7 (GWi7.1/ GWS.7,1, GWS7.2/GWi7.2) for both GWi and GWS11.1 genetic region associated with only GWi. Detailed inferences on chromosome 3, 5 and 7 were presented together with indel results (see next sections).

A GWAS peak at 28989431–29000913 bp (size = 11.5 kb) on chromosome 11 identified two novel SNPs that have highly significant associations with GWi in japonica but not for GWi (Fig. 1). These two SNPs, snp11_28989509 (log10(P) = 9.3) and snp11_28989516 (log10(P) = 11.36), had effect of 0.94 and 1.02, respectively (Tables S1 and S2). These two SNPs were located at the 3′-UTR of LOC_Os11g48090 that was known to encode for helicase conserved C-terminal domain containing protein (Fig. 4). Two alternative haplotypes (Fig. 4B) formed by these SNPs: haplotype 1 (TC) was observed in 134 genotypes with mean GWi of ~2.9 mm, while haplotype 2 was heterozygous (T/C, C/T) and was found in 124 genotypes with mean GWi of ~2.5 mm. Mapping these two haplotypes to the 3,000 Rice Genomes showed that haplotype 1 was prominent in temperate japonica and aromatic and at moderate levels in indica. It was under-represented in aus and tropical japonica (<18%) (Fig. 4B). Haplotype 2 appeared to have more representation in aus, and tropical japonica germplasm,
Figure 1. SNP-based GWAS for grain width (GWi) that confirmed GW5 and identified prominent candidate loci on chromosomes 7 and 11. Manhattan plots of the genome-wide association studies on GWi for japonica (A), indica (B) and all (C, combined) panels (left side) for raw grain; japonica (D), indica (E) and all (F, combined) for cooked grain (right side). The novel genomic regions in chromosomes 7 (GWi7.1, GWi7.2, GWi7.1, GWi7.2) and 11 (GWi11.1) were detected along with previously characterized/cloned gene in chromosome 5 (GW5). Horizontal red and blue line represents the genome-wide significant threshold $−\log_{10}(P)$ value of 7 and 5, respectively.
Figure 2. SNP-based GWAS for grain length (GL) that confirmed GS3. Manhattan plots of the genome-wide association studies on GL for *japonica* (A,D), *indica* (B,E) and combined (C,F) panel in raw and cooked grain, respectively. The topmost significant SNP (snp_03_16733441) in both raw and cooked grains was detected within the genic region of GS3. The boxplot (mentioned in B,C,E and F) visualized phenotypic variations in raw grain length (B,C) in *indica* and cooked grain length (E,F) in the combined *indica* and *japonica* panel for each haplotype formed by two of the most significant SNPs. The haplotype with the A allele of the most significant SNP (C to A) had the highest contribution to the grain length variation in both raw and cooked grains. Horizontal red and blue line represents the genome-wide significant threshold $-\log_{10}(P)$ value of 7 and 5, respectively.
Figure 3. SNP-based GWAS for grain shape (GS) and association signals detected on chromosome 3, 5, 7, 10 and 11. GS being a derived trait from the ratio of GL to GWi, association signals detected in both GL and GWi were expected to show up in GS. Results showed that the cloned genes for GWi (GW5) and GL (GS3) were detected, and new candidate loci in chromosomes 3, 7 and 11 detected previously in GW and GL GWAS were also detected in japonica (A,D), indica (B,E) and all (C,F combined) panel (left side) in raw (left) and cooked grain (right), respectively. A new locus in chromosome 10 (GS10.1) was also detected. Horizontal red and blue line represents the genome-wide significant threshold $-\log_{10}(P)$ value of 7 and 5, respectively.
moderate representation in indica and to a lower frequency in aromatic and temperate japonica (<30%) (Fig. 4B).
Interestingly, the homozygous rare haplotype (CT) was found in extremely low percentage in tropical japonica (Fig. 4B).

Novel region detected by GWAS such as GS10.1 from chromosome 10 associated with cooked grain shape detected only in SL-GWAS method with higher significance, where a single prominent significant SNP (C→T; effect = −0.58, −log10(p) = 9.83) was detected in the intronic region of gene E3 ubiquitin ligase (LOC_Os10g35920). Mining the potentially causal variant in 3 K RGP data suggest a limited presence of the T allele within tropical japonica and Japx (higher ratio of GLc/GWi c) in cooked rice grains (mean 3.48) than rest of the germplasm with C allele (mean GS 2.89) (Fig. 4A).

Figure 4. Gene structures of LOC_Os10g35920 and LOC_Os11g48090, the phenotypic variation explained by their haplotypes, and penetrance in the 3,000 rice genomes. (A) LOC_Os10g35920 was implicated by a single intronic SNP (GS10.1) that had a significant association to cooked grain shape (GS1). The T allele of this SNP had the most influence in the GS1 variation that when combined with the G alleles of the two promoter region SNPs has the potential to result in bolder grains. Mapping haplotype TGG in the 3,000 rice genomes showed that the haplotype was present only in japonica accessions, particularly tropical japonica. (B) The gene structure of LOC_Os11g48090 showed the position of the significantly associated SNPs positioned at the 3′-UTR of the gene. Plotting the variation of GWi with respect to the haplotypes formed by the significant SNPs showed that the accessions that were homozygous to the T and C alleles of these SNPs showed a tendency to be bolder both in raw and cooked grains. Mapping the haplotypes in the 3,000 rice genomes showed that all haplotypes have pervasive representations in all subspecies as well as those classified as admixtures.

Indel-based GWAS results of grain size and shape traits. A complementary GWAS analysis that used 393,429 indels highlighted associations on chromosome 3 for GL, and chromosomes 5 and 7 for GWi (Figs S5–S7). These results matched the SNP-based GWAS on their respective regions with SL (Figs 1–3) and ML approaches (Fig. S8).

The indel-based GWAS analysis revealed a 0.28 Mb region (16.90 Mb–17.20 Mb) with high linkage disequilibrium (LD) decay in chromosome 5 downstream from GS3 that associated with GL in indica (Fig. 5A). Further analysis showed that haplotypes formed by indels in two genes (LOC_Os03g29710 and LOC_Os03g29730) were able to explain certain ranges in the phenotypic variation of GL (Fig. 5B,C). Interestingly, none of the significant novel allelic variants within GS3 that explains significant associations were detected by indel-based GWAS analysis. A highly significant SNP-based GWAS association signal (−log10(p) ≥ 17) for GLc and GL were observed in a 0.42 Mb region (16.66 Mb – 17.11 Mb) on indica chromosome 3 (Fig. S9). Mining the potential haplotypes, and penetrance in the 3,000 rice genomes. (A) The gene structure of LOC_Os10g35920 showed the functional polymorphism present in the region between the LOC_Os05g09510 and LOC_Os05g09520, was a 1,212 bp deletion in the 3′-UTR of the gene. Plotting the variation of GWi with respect to the haplotypes formed by the significant SNPs showed that the accessions that were homozygous to the T and C alleles of these SNPs showed a tendency to be bolder both in raw and cooked grains. Mapping the haplotypes in the 3,000 rice genomes showed that all haplotypes have pervasive representations in all subspecies as well as those classified as admixtures.

Haplotypes formed by novel causal variants on chromosome 7 explain variation in raw and cooked grain widths. Through SNP-based ML- and SL-GWAS analysis, prominent association signals were detected on chromosome 7 for GWi and GWi c when both of the subspecies (indica and japonica) were combined. This reflected the contribution of large effect allelic combination contributed by both subspecies
Interestingly, less prominent association signals for GWi ($-\log_{10}(p) \geq 5$) were detected on chromosome 7 within subspecies GW AS (Fig. 1A,B,D,E). Within this broad peak on chromosome 7, two different regions [(GWi7.1/GWi c7.1 at intervals 22.1 Mb–22.8 Mb) and (GWi7.2/GWi c7.2 at interval of 23.3 Mb–25.2 Mb)], were identified to influence GWi and GWi c (Fig. S11). Indel-based GW AS analysis identified 21.84 Mb–23.05 Mb that overlapped GWi7.1/GWi c7.1 to associate with GWi; no signal was detected at GWi7.2/GWi c7.2.

A total of 24 tag SNPs at GWi7.1/GWi c7.1 formed two LD blocks with three unique haplotypes in each block (Fig. 6A,B, Fig. S11, Table S2). SNPs 3, 7, 15, and 16 had the strongest association likelihood and the largest allele effects among the SNPs. The indel-based GW AS results identified 12 significant structural variants with three LD blocks within GWi7.1/GWi c7.1 (Fig. 6C,D). The GW AS peak at this region was dissected using PLINK’s –clump function. Indels in block 1 and 2 showed larger allele effects compared to those indels in block 3. Indel-based targeted association of candidates within GWi7.1/GWi c7.1 identified 4 candidate genes (LOC_Os07g37150, LOC_Os07g37156, LOC_Os07g37820 and LOC_Os07g37920) with indels and also confirmed to possess SNP-based haplotypes that can be used to discriminate GWi. Two additional non-synonymous SNPs and an indel in 3′-UTR detected in LOC_Os07g37920 (NAC transcription factor) were found to associate with GWi (snp_07_22685420; A $\rightarrow$ G) alter the protein sequence by replacing charged arginine (R) with hydrophobic glycine (G), found to associate with raw (Effect = 0.31, $-\log_{10}(p) = 6.65$) and cooked grain width. In addition, 10 other SNPs in untranslated and intronic regions were found in LOC_Os07g37820 (Fig. 7A, Table S1). A total of 490 accessions that had the haplotype TATTCGCTCA had means of 2.6 mm and 3.8 mm for GWi and GWi c, respectively. A total of 59 accessions that had haplotype TATTCGCTCA had means of 2.2 mm and 3.3 mm for GWi and GWi c, respectively (Fig. 7A).

Using the SL-GW AS approach, targeted gene association analysis in GWi7.1/GWi c7.1 region identified six additional loci (Fig. S13). An SNP (snp_07_22608875) lying in downstream region of the gene LOC_Os07g37710...
Figure 6. GWAS for grain width (GWi) using SNPs and indels revealed the association of a region in chromosome 7 to GWi. (A) The linkage disequilibrium (LD) plot of the 24 tag SNPs significantly associated with grain width. A scaled and highly dense plot of the associated genomic region on the chromosome is shown where the relevant genes are marked in red (boxes). The positions of the 24 tagged SNPs are also marked with the log_{10}-scaled association P values of these 24 SNPs are shown in the bar plot where black bars reflect their relative effect sizes. The gene IDs further detected in T GAS were highlighted in red color. (B) Haplotypes constructed based on SNPs in LD are represented as boxplot with the phenotype values for both normal and cooked grain explained by specific haplotype. Also shown are (C) the linkage disequilibrium plot for indels associations for the chromosome 7 with black bar graph signifies effect size on the grain width and (D) Haplotype constructed with phenotype distribution within each blocks formed from the significant indels in the region represented as boxplot for both cooked and raw grain.
detected to significantly affect GWi \( (\beta = 0.44, -\log_{10}(p) = 8.72) \) (Fig. S13A, Table S2). Low GWi haplotype TTA represented the mean GWi and GWi c of 2.1 and 3.3 mm, respectively and mainly detected in tropical japonica, Japx and aus cultivars upon scanning 3000 rice germplasm. Another SNP (snp_07_22662700) present on intronic region of LOC_Os07g37790 (transcription regulator), positively associated with GWi \( (\beta = 0.51, -\log_{10}(p) = 9.37) \) (Fig. S13B). Similarly, low GWI haplotype T showed the mean GWI and GWI c of 2.1 and 3.3 mm, respectively and detected in tropical japonica, Japx and aus. In LOC_Os07g36900 (F-box domain protein) a total of seven SNPs including three non-synonymous SNPs (snp_07_22097823, snp_07_22097824 and snp_07_22097863) were identified (Fig. S13C). These adjacent non-synonymous SNPs in exon 1 were annotated to alter the amino acid sequence by replacing charged aspartate with hydrophobic alanine. Within LOC_Os07g36900, haplotype AACATTT was found in 34 genotypes with mean GWi of 2.1 mm and GWi c of 3.1 mm. Haplotype GGTGGCA was found in 508 lines with a relatively higher mean GWI and GWI c of 2.6 and 3.8 mm, respectively (Fig. S13C). The identified genes associated with GWi were evaluated for their expression in low to high grain width genotypes and identified a candidate LOC_Os07g36900, differentially expressing across the lines possessing contrasting phenotypes (Fig. S13). Additionally, the moderate to high expression of gene in later growth and development stages specially in heading stage suggest their role in grain size determination (Fig. S15).

SNP based GWi7.2/GWi c7.2 region had three LD blocks and matched a known gene GL7/GW7 (LOC_Os07g41200) (Fig. S12). By employing the ML- and SL-GWAS approaches, three novel candidates (LOC_Os07g38980, LOC_Os07g41000 and LOC_Os07g41080) were identified to associate with GWI and GS in target region GWI 7.2 (Fig. 7, Table S1). Within LOC_Os07g38980, GACCCGAAT haplotype was identified for fixing and retaining the lower GWI in tropical japonica/japx and aromatic lines (Fig. 7B,C). Upon TGAS, intronic-SNP (snp_07_23368244; C→T) detected in novel gene LOC_Os07g38980 encoding unclassified protein significantly associated with raw and cooked GWI (Fig. 7B,C). Throughout the genome ML-GWAS independently yielded a number of candidates associated with cooked and raw grain dimension traits (Table S3). In addition, ML-GWAS also identified and confirmed the loci detected with less significant critical value using SL-GWAS, owing to stringent correction criterion (Table S3).

---

**Figure 7.** Targeted-gene association study (TGAS) in hotspot region of Chr7 (GWi7) commonly detected in single- and multi-locus GWAS identified a total of five genes that had significant association with GWI and GWI c. (A) LOC_Os07g37820 was identified as key gene in GWI 7.1 region. Three additional genes (B) LOC_Os07g38980, (D) LOC_Os07g41000 and (E) LOC_Os07g41080 were identified through TGAS in GWI 7.2 region. The blue inverted triangles in the gene structure diagram represent the SNPs included in plotting the boxplots. Boxplots were generated to show the phenotypic distribution of haplotypes for raw and cooked GWI. Respective haplotypes mapped in 3,000 rice genome results were shown on the right panel.
Mapping grain size and shape genes to synteny map created between *japonica* and *indica* reference genomes. In order to dissect the degree of structural variation, the synteny relationship at subspecies level was evaluated by doing comparative genomics between Nipponbare (MSU Release 7) and 93-11, reference genomes of *japonica* and *indica*, respectively. We found that synteny was high for chromosomes 3, 5 and 7 with apparently several random break points occurring throughout these chromosomes (Fig. 8). Chromosome 11 depicted severe erosion in collinearity. We have noted significant SNP- and indel-based GWAS association signals on chromosomes 3, 5 and 11 that occurred at syntenic break points (Fig. 8). GL3.1, a region that has an uncharacterized gene at a syntenic break point in chromosome 3, was detected by indel-based GWAS to be significantly associated with GL. For this trait, the most significantly associated SNP detected by a SNP-based GWAS was from the cloned G35 gene found at the conserved region of chromosome 3 of *indica* and *japonica*. For GWi and GW1, we found an association signal at GWi5.1 on chromosome 5 that confirmed the effect of the deletion upstream of GW5 in *japonica* that is absent in *indica*. GW5 was implicated by several significantly associated SNPs at its promoter region. It is interesting to find neighborhood of major GW regulating gene GW7 at the collinearity break between sections of chromosomes 5 in both *indica* and *japonica*. We also similarly fine-mapped GW11.1 that matched to a collinearity break point in chromosome 11 significantly associated with grain size and shape (Fig. 8). Significant association signals were exclusively detected on chromosome 11 in *japonica*. The break points in chromosomes 3, 5 and 11 of *japonica* were due to structural variations such as insertion or deletion (Fig. 8). These structural variations at the subspecies level led to the loss of collinear regions within *indica* and *japonica* subspecies.

It is interesting to note that many of the cloned genes for grain size in rice were found to match the collinear regions in the syntenic map (Fig. 8). TGAS for known genes which were previously cloned for grain dimensions and shape enabled the detection of additional alleles and confirmed previously reported SNP variants, though the respective association likelihoods were weak to moderate (Table S4, Fig. S14). The allelic variations at the subspecies level that lie within *Gif1*, GL3.1, *big grain2*, and *short grain1*, genes that were known to regulate grain size and shape in both *indica* and *japonica*, could likely be attributed to preferential selection of alleles in both subspecies during domestication. The allelic variations in *An-1*, *GS5, small grain11* were preferentially enriched in *japonica*, while GL7/GW7, *small grain1, srs5* allelic variation were enriched in *indica* subspecies (Table S4, Fig. S14).

Discussion

Grain dimensions are major determinants of grain weight and therefore are important component traits affecting crop yield. These traits also influence varietal acceptability to consumers, and therefore rice grain size/shape is a major preferential target trait in breeding. Short and bold type cultivars are highly preferred by many consumers in Japan, South Korea and northern China, whereas consumers in India, the USA, and other South and Southeast Asian countries favor long slender and medium slender grains. Many of the genes/QTLs that regulate grain length, width and thickness function in selective proteolysis, as well as those that promote cell proliferation and expansion have been cloned, characterized and further validated in populations developed either from crossing within or across the subspecies (Table S4). These traits have been shown to be controlled by both major and minor genes or QTLs that often exhibit additive effects, dominant effects or both.

Previously, moderate to high coverage SNP genotype data generated through genotyping by sequencing and Affymetrix array-based genotyping platforms were used to perform GWAS analysis and to decipher the genetics of rice grain length. In the present study, high-density genotype data derived from re-sequenced genomes of 591 diverse landraces were used to identify the novel allelic variants and the haplotypes that underlie grain length, width and shape in raw and cooked rice grain. A total of ~2.9 million high-quality SNPs and 393,429 indels were used as the genotype discovery resource to perform SNP- and indel-based GWAS analyses. These genotype data densities and the substantial level of population divergence and within-population genetic variation evident in *indica* and *japonica* make it highly likely for GWAS to catch trait-associated alleles. The resolution provided by the dense SNP markers was at least three-folds higher than the recently published reports on GWAS in cultivated rice. This clearly reflects the advantage of high-coverage re-sequencing based SNP genotyping resources in the dense SNP markers was at least three-folds higher than the recently published reports on GWAS in cultivated rice. Affymetrix array-based genotyping platforms were used to perform GWAS analysis and to decipher the genetics of rice grain length, width and shape in raw and cooked rice grain. A total of ~2.9 million high-quality SNPs and 393,429 indels were used as the genotype discovery resource to perform SNP- and indel-based GWAS analyses. These genotype data densities and the substantial level of population divergence and within-population genetic variation evident in *indica* and *japonica* make it highly likely for GWAS to catch trait-associated alleles. The resolution provided by the dense SNP markers was at least three-folds higher than the recently published reports on GWAS in cultivated rice. This clearly reflects the advantage of high-coverage re-sequencing based SNP genotyping resources in the dense SNP markers was at least three-folds higher than the recently published reports on GWAS in cultivated rice. In order to dissect the degree of structural variation, the synteny relationship at subspecies level was evaluated by doing comparative genomics between Nipponbare (MSU Release 7) and 93-11, reference genomes of *japonica* and *indica*, respectively. We found that synteny was high for chromosomes 3, 5 and 7 with apparently several random break points occurring throughout these chromosomes (Fig. 8). Chromosome 11 depicted severe erosion in collinearity. We have noted significant SNP- and indel-based GWAS association signals on chromosomes 3, 5 and 11 that occurred at syntenic break points (Fig. 8). GL3.1, a region that has an uncharacterized gene at a syntenic break point in chromosome 3, was detected by indel-based GWAS to be significantly associated with GL. For this trait, the most significantly associated SNP detected by a SNP-based GWAS was from the cloned G35 gene found at the conserved region of chromosome 3 of *indica* and *japonica*. For GWi and GW1, we found an association signal at GWi5.1 on chromosome 5 that confirmed the effect of the deletion upstream of GW5 in *japonica* that is absent in *indica*. GW5 was implicated by several significantly associated SNPs at its promoter region. It is interesting to find neighborhood of major GW regulating gene GW7 at the collinearity break between sections of chromosomes 5 in both *indica* and *japonica*. We also similarly fine-mapped GW11.1 that matched to a collinearity break point in chromosome 11 significantly associated with grain size and shape (Fig. 8). Significant association signals were exclusively detected on chromosome 11 in *japonica*. The break points in chromosomes 3, 5 and 11 of *japonica* were due to structural variations such as insertion or deletion (Fig. 8). These structural variations at the subspecies level led to the loss of collinear regions within *indica* and *japonica* subspecies.

This study enabled the discovery of large numbers of SNPs and indels that differentiate the important traits especially grain width and shape (Fig. 8). Significant association signals were exclusively detected on chromosome 11 in *japonica*. The break points in chromosomes 3, 5 and 11 of *japonica* were due to structural variations such as insertion or deletion (Fig. 8). These structural variations at the subspecies level led to the loss of collinear regions within *indica* and *japonica* subspecies.

It is interesting to note that many of the cloned genes for grain size in rice were found to match the collinear regions in the syntenic map (Fig. 8). TGAS for known genes which were previously cloned for grain dimensions and shape enabled the detection of additional alleles and confirmed previously reported SNP variants, though the respective association likelihoods were weak to moderate (Table S4, Fig. S14). The allelic variations at the subspecies level that lie within *Gif1*, GL3.1, *big grain2*, and *short grain1*, genes that were known to regulate grain size and shape in both *indica* and *japonica*, could likely be attributed to preferential selection of alleles in both subspecies during domestication. The allelic variations in *An-1*, *GS5, small grain11* were preferentially enriched in *japonica*, while GL7/GW7, *small grain1, srs5* allelic variation were enriched in *indica* subspecies (Table S4, Fig. S14).

Previously, moderate to high coverage SNP genotype data generated through genotyping by sequencing and Affymetrix array-based genotyping platforms were used to perform GWAS analysis and to decipher the genetics of rice grain length. In the present study, high-density genotype data derived from re-sequenced genomes of 591 diverse landraces were used to identify the novel allelic variants and the haplotypes that underlie grain length, width and shape in raw and cooked rice grain. A total of ~2.9 million high-quality SNPs and 393,429 indels were used as the genotype discovery resource to perform SNP- and indel-based GWAS analyses. These genotype data densities and the substantial level of population divergence and within-population genetic variation evident in *indica* and *japonica* make it highly likely for GWAS to catch trait-associated alleles. The resolution provided by the dense SNP markers was at least three-folds higher than the recently published reports on GWAS in cultivated rice. This clearly reflects the advantage of high-coverage re-sequencing based SNP genotyping resources in *indica* population where LD decays faster (Supplementary Note) compared to *tropical* and *temperate japonica*. This study enabled the discovery of large numbers of SNPs and indels that differentiate the important traits especially cooked rice grain dimensions and shape that are being selected during domestication. For cooked grain, a distinct allelic variant was revealed in the intron region of gene encoding E3 ubiquitin ligase on chromosome 10, exerting substantial effect on cooked grain shape. Other E3-ubiquitin ligases were earlier revealed to regulate the grain width in rice and grain size in other crops. GWAS analysis performed on a genome-wide SNP matrix that was designed to cover coding as well as the regulatory regions such as promoters, alternative spliced junctions, 3′ and 3′ untranslated region helped in the discovery of novel regulatory alleles and their haplotypes. Genetic dissection of the genomic region detected on chromosome 11 in *japonica* revealed the presence of two adjacent SNPs (7-bp apart) in the 3′-UTR region of LOC_Os11g48090, a gene that encodes helicase conserved C-terminal domain containing protein in rice (Fig. 4C). These SNPs showed strong associations with grain width. Interestingly, none of the other SNPs in this region were implicated by LD, to influence grain width. In addition, a short region containing causal SNPs present within 3′-UTR showed ~90% homology with the binding site of orthologous members of High Mobility Group (HMG) Box DNA-binding proteins (namely AHL12 and AHL25), previously known to have a role in Arabidopsis growth and development. Transcriptional repressors target the 3′-UTRs and the region near the stop codon. Therefore, nucleotide substitutions noted in the 3′-UTR region of gene LOC_Os11g48090 potentially lead to the disruption of the binding site of a repressor. This region requires further functional validation. The haplotypes comprised of these causal SNPs that are responsible for conferring wider grain width (mean = 2.9 mm), were detected in *temperate japonica*, and heterotic alleles T/C-C/T explaining lesser grain width (mean = 2.5 mm) abundantly.
Figure 8. Detailed sequence collinearity in chromosomes 3, 5, 7, 10 and 11 between indica and japonica reference genomes. Circos plots showed syntenic relationship of five chromosomes (physical size shown in Mb) within indica and japonica subspecies based on protein sequence alignment. Conserved regions were shown in green while black lines represented the break points. The positions of genes that have been cloned for grain width, length and shape were also labeled in the synteny map. Blue and green boxes represented regions in the respective chromosome where GWAS peaks were detected; blue were for SNP-based GWAS, while green were for those done using indels. It was evident that most of the cloned/characterized genes for grain size and shape were in collinearity between indica and japonica.
Genomic analysis on global MAGIC populations, developed specifically to explore effects of allelic recombination between *japonica* and *indica* subspecies, was able to identify a region located at position 22.7 Mb – 26 Mb on chromosome 7 that significantly associates with grain width. GWAS in this study performed in a combined *indica* and *japonica* populations was able to identify GWAS peaks as GW7.1 and GW7.2 on chromosome 7 which have significant distance to affect independently the grain width. This region overlaps with what was previously reported for raw using the MAGIC population and bi-parental population. This also suggests that functional allelic variations present in the two major subspecies were highlighted effectively when combined. Our study further narrowed down this major effect region into two sub-regions with GW7.1 mapped at the interval of 22.1–22.8 Mb and GW7.2 narrowed down to a region of 23.3–25.2 Mb. Employing integrated multi-locus and single-locus GWAS approach (EMMAX) led to verify the significance of underlying target regions, GW7.1 and GW7.2 and simultaneously identify novel candidate genes. Furthermore, TGSAS of the GW7.1 region identified four putative candidate genes that are significantly associated with grain width. Among these, one encodes for NAC-transcription factor and three are unknown protein encoding genes that possess indel haplotypes and non-synonymous SNPs differentiating cooked and raw grain width phenotypes (Fig. 7, Fig. S13). Identification of NAC transcription factors as candidate genes is coherent with the reported role possessing indel haplotypes and non-synonymous SNPs differentiating cooked and raw grain width phenotypes. Among these, one encodes for NAC-transcription factor and three are unknown protein encoding genes that were representing novel variations were spotted along with previously characterized loci within GL7/GW7 that were known to regulate grain length/width and identified 3 additional genes (LOC_Os07g38980, LOC_Os07g41000, and LOC_Os07g41080) influencing grain width. Employed multi-locus GWAS was helpful to validate novel loci identified from the genomic regions using single-locus GWAS method with a less stringent significance criterion. The present study demonstrated the power and resolution of whole genome re-sequencing to identify novel genomic variants and their haplotypes that contribute significantly in providing and fixing narrower grain width for both raw and cooked rice grains (~2.0 mm/~3.0 mm respectively).

Many of the cloned size and shape related genes comprising GW8/OsSPL16, SR5 and TGW6 were mapped in conserved regions. From high-resolution SNP- and indel-based GWAS, several significant associations that mapped to synteny breaks were revealed in this study. These findings show how structural variation influences grain size traits. In the *indica* population, the GL3.1 association signals for grain length on chromosome 3 (16.12–16.40 Mb) and GW5.1 detected on chromosome 5 (5.36–5.47 Mb) grain width were mapped to the break points. Comparative genomics to study synteny in this region unveiled extensive sequence collinearity with intermitten gaps within gene contents, intergenic regions and gene orders across *indica* and *japonica* reference genomes. This was consistent with the previous reports on rice genome diversity. The breakage in synteny was confirmed due to several significant SNP variation and deletion detected upstream of cloned GW5 region influencing grain size/shape. The rearrangements leading to insertions and deletions caused larger synteny breakage relative to others, for instance, in case of upstream of GW5, a 1212 bp region was deleted in Nipponbare genome (*japonica*) corresponding to ~2300 bp fragment of Kasalath genome (*indica*) . Although several important alleles explaining major QTLs for grain dimensions and shape in raw rice grains were confirmed in *indica* and GW5 genes, this study has identified additional causal indels particularly in GW3.1. Additionally, the GWAS peak identified for GW11.1 found only in *japonica* overlapped a break region.

Using structural variants as genotype data in GWAS offered new insights to map novel alleles located in the break regions that influence grain size and shape. The use of *japonica* reference genome in this study to account for indel variation across *japonica* and *indica* sub-species was informative and effective in identifying novel genes located at regions where the synteny between these two subspecies breaks. For future studies, taking the subspecies level structural variations into account by mapping the reference genomes and calling SNPs using a much improved *indica* reference genome could further enhance the precision and resolution of the genome-wide mapping studies at the subspecies level. In this regard, utilization of the recently released map-based high-quality genome sequence and annotations of major varieties (Zhenshan 97 and Minghui 63) of *indica* as a reference could potentially lead to even more profound understanding of the genetics of certain traits.

A novel approach implemented in the present study, where the result of comparative genomics between reference genomes of two major *Oryza sativa* L. sub-species used in conjunction with GWAS using whole genome re-sequencing resources to identify structural variation in genes influencing cooked and raw grain size and shape in rice. Through this pipeline: (1) Significant association signals that implicated novel genes associated with cooked and raw grain width, length and shape for each subspecies were identified through SNP- and indel-based GWAS analysis; (2) Diagnostic haplotypes were defined through LD analysis and tag SNPs approaches; (3) The syntenic maps were derived between the subspecies and mapped the GWAS peaks to identify genetic regions which falls in collinear break points, and conducted targeted associations to reveal novel gene based haplotypes; (4) Mining the 3000 Rice Genomes for the rare haplotypes found in this study revealed rare germplasm that fall into specific grain width and length ranges in each *Oryza sativa* L. sub-species. This study has built a series of genomic pipelines that expedited the identification of novel grain dimensions and shape genes and also provided a comprehensive understanding of the distribution of alleles and haplotypes that are deemed preferentially selected for grain dimensions and shape for both raw and cooked grains though the process of domestication.

**Methods**

**Plant materials.** A total of 591 diverse accessions whose days to maturity do not exceed 140 were selected from the 3000 resequenced rice genomes (Table S5). This panel was comprised of 324 *indica* and 267 *japonica*. They were grown in a one and a half hectare contiguous experimental area at the International Rice Research
Phenotyping for measurement of grain dimensions. Quantities of 50 g seed material from every line with independent replicates were obtained and subsequently equilibrated at room temperature before any physical analysis was done. For cooking, rice grains were cooked at 100°C using the standard operating procedure of IRRI Grain Quality and Nutrition Service Laboratory (GQNSL). Grain size and shape were measured using the ISO 17025 certified protocols of the IRRI GQNSL that used a SeedCount SC5000 Image Analyser (http://www.knowledgebank.irri.org/ricebreedingcourse/body.default.htm#Grain_quality.htm).

Genome-wide SNP identification. A total of 591 diverse landrace accessions composed of 324 indica and 267 japonica varieties were used for the genome-wide SNP identification. Publicly available variant call format (VCF) files from the published 3,000 rice genomes were used to compile the SNPs. These VCF files contained the complete base calls across the entire genome from which SNPs and INDELS were identified. Each VCF file was filtered using VCFtools to keep only those base calls with a Phred score of 30 or better. After filtering, the quality assured VCF files were then merged, all SNP and INDELS were then extracted separately, saved into two different VCF files for SNPs and INDELS, and then subsequently converted into PLINK format (BIM, BED, FAM).

Population structure and calculation of linkage disequilibrium decay. Principal components analysis (PCA) was performed in diversity lines using the SNPRelate package in R to detect the population structure. An LD cut-off of 0.99 was used that resulted in the selection of 673,846 SNPs in PCA calculation. The first two principal components accounted for 42.57% of the total genetic variation. For linkage disequilibrium (LD) decay prediction, we calculated the pairwise LD of all SNPs present in the three different populations using PLINK v1.90 beta (PLINK2). Then, bins representing multiples of 50-kb distances between SNPs were formed where the mean r² for each bin calculated using a custom PERL script and plotted. This was done for each sub-population.

GWAS analysis. The genotype file was filtered prior to running GWAS analysis. Using plink2, we retained individuals and SNPs that had a missing rate of not more than 5% and then filtered for a minor allele frequency of at least 5%. This filtering step resulted to a final set of 2,933,037 SNPs and 585 distinct varieties from both indica and japonica subspecies. japonica accessions accounted for 1,562,079 total numbers of SNPs from 267 individuals, while 324 indica lines accounted for 2,260,030 SNPs. A total of 393,429 indels were considered after filtering with missing rate of not more than 5% and then filtered for a minor allele frequency of at least 5%. Phenotype data was then transformed using WarpedLMM to satisfy the data distribution requirement of mixed linear model separately for SNPs and INDELS. WarpedLMM uses a monotonic warping function to transform the phenotype data where instead of using a static function it searches for a most suitable transformation function for the given phenotype data. EMMAX was used for computing the single-locus association statistics where the kinship and population structure were added as covariates into the mixed linear model for both SNPs and INDELS. The statistical model underneath the EMMAX uses variance component model as described in the Kang et al. (See Online Methods of the paper for the complete details) that belongs to the family of mixed linear models. This approach to GWAS effectively corrects the confounding factors such as relatedness between samples and population structure that if uncorrected would lead to spurious associations. In this model, the predictors were composed of 1.5, 2.2 and 2.9 million SNPs for japonica, indica and combined population, respectively along with the identity-by-state kinship matrix and the population structure computed using principal components analysis. The markers were considered as fixed effects, while the kinship matrix and principal components were treated as random effects. We assumed that each term is normally distributed, although EMMAX did not explicitly mention the requirement. Beta-coefficient, indicating the effect size of the marker on respective phenotype, was mentioned as ‘effect’ in the outputs of GWAS.

Kinship was calculated using emmax-kin while the population structure was represented using the first two principal components. IBS matrix was used as kinship matrix. Two principal components were sufficient covariates for the combined indica and japonica set, as well as the indica-only set, while three principal components were required for the japonica set. These decisions were based on the scree-plot derived from the PCA results. The statistical model used for GWAS was a variance component model reported in the Kang et al. The threshold value was set at 1.70e-08 using Bonferroni correction (or \(-\log_{10}(0.05/2933037) = 7.77\) as shown in the Manhattan plot) for identifying the peak association signals, however, an LD-based tagged SNP criteria was followed on SNPs with \(-\log_{10}P > 5\) for detailed analysis on raw grain phenotype. The complete parameters when clumping for SNPs and INDELS in plink2 were \(-\text{clump-p1 1e-7, } -\text{clump-p2 1e-5, } -\text{clump-kb 200, } -\text{clump-r2 0.5}\). These parameters ensures that the “index” SNPs around which the “clumps” were formed must have a p-value of at most 1e-7, while those SNPs forming the clump around the index SNP must have a p-value of at most 1e-5. Due to less significant SNPs–clump-p2 parameter was considered as 0.01 (p-value) in case of cooks grain width by following the clumping method outlined in our previous study. The farthest SNP that may be clumped to
the index SNP was 200 kb with an LD $r^2 = 0.5$. All genomic positions and gene annotations were based on the Nipponbare reference genome (MSUv7). Tag SNPs were identified using Haploviz for groups having an LD coefficient $D' \geq 0.8$, while haplotype blocks were also formed from the same result$^{46}$. In-house R-scripts were used for creating $-\log_{10}(p)$ value plots and box plots for depicting phenotype distribution within the designated haplotypes. Targeted-association study was performed where only genic SNPs and those falling within 2 kb upstream and 1 kb downstream of the genomic region. After TGAS we plotted the significant SNPs showing causal association with phenotype ($\log_{10}(p)$ value). Non-synonymous SNPs were determined upon completing SNP annotation using annovar.

**Multi-locus GWAS analysis.** We reanalyzed the GWAS on a same set of population earlier conducted through single-locus (SL)-GWAS, using three different multi-locus (ML) GWAS tests–FASTrEMMA$^{25}$, mRMLM$^{24}$ and ISIS EM-BLASSO$^{23}$. SNP pruning was performed on the entire SNP set (mentioned above) since memory constraint was observed with entire SNP set to run MLA tools. Therefore, a window size of 5 Kb (to include more SNPs in the analysis), a step size of 5 SNPs in each step and $r^2 = 0.5$ were used as the pruning parameters in PLINK. Finally, with the 1393842 SNPs (combined indica and japonica), 1079207 unique SNPs for indica and 586697 unique SNPs for japonica were extracted and were directly used for conducting ML-GWAS using all of the three aforementioned methods. Default critical $p$-value criterion was adopted as per the details mentioned in respective method. Two principal components were used for both indica specific GWAS and for the combined set of indica and japonica, whereas three principal components were used for japonica population. Manhattan plots were created using the first step result of the multi-locus association. For tabulation of the SNP loci from all of the three methods, LOD $\geq 3$ were considered as threshold parameter. The genomic regions surpassing the threshold significance criteria of LOD $\geq 3$ (in case of ML-GWAS) and $-\log_{10}p \geq 5$ were considered as common regions between ML- and SL-GWAS methods. The genetic regions simultaneously and individually identified in ML- and SL-GWAS were further categorized in respective Tables S1–S3.

**Transcriptome/expressional analysis.** The expression profiling across different stages of plant growth and development was determined by using Affymetrix rice genome array at public database Genevestigator$^{62}$. The log2 transform values were utilized to construct a plot. For differential expression profiling, total RNA from selected lines was extracted from developing (16 d post anthesis, dpa) grains for transcriptome analyses utilizing a genome-wide microarray platform (Agilent Technologies) (methods adapted from$^{65}$). The gene expression profiling was conducted by hybridizing onto a genome wide microarray slide for rice based on the manufacturer’s protocols (Agilent Single Color; Agilent Technologies). The data was normalized using GeneSpring GX (Agilent, Santa Clara, CA) following quantile normalization algorithm. Lines with contrasting phenotypes were selected for expression analysis and the log2 transform value has been shown as heat map using Genesis tool$^{63}$ with variance of $\pm 3$.

**Mapping out the haplotype blocks to the rest of the 3,000 Rice Genomes.** SNP-seek database$^{64}$ was used to determine the enrichment of the phenotype discriminating haplotype blocks in each of the different subspecies represented in the complete 3,000 rice genomes panel. The distribution representation of the haplotype blocks in each subspecies was calculated and visualized as percentages.

**Synteny overlays with grain size and shape genetic regions.** We used reference protein sequences of japonica and indica, adapted from reference genomes of MSU version 7 and gramene database, respectively. We followed the all-to-all blastP of respective protein sequence using NCBI-BLAST-2.2.28$^{+}$ tool$^{65}$. A stringent criterion of e-value of e-30 was used in blast alignment of protein sequence within japonica and indica considering their genetic similarity at sub-species level. Subsequently, collinearity was established using MCScanX$^{66}$ with the threshold of 10 genes constructing each collinear block. On the basis of these outcomes, collinearity was identified and represented in the form of circos$^{67}$ that further overlaid with genomic region corresponding to cloned and characterized genes regulating grain size and shape and significant genomic region regulating grain width detected as an outcome of GWAS.

**References**

1. Kovach, M. J., Sweeney, M. T. & McCouch, S. R. New insights into the history of rice domestication. *Trends Genet.*, 23, 578–587 (2007).
2. Sweeney, M. & McCouch, S. The complex history of the domestication of rice. *Ann. Bot.*, 100, 951–957, https://doi.org/10.1093/aob/mcm128 (2007).
3. Famoso, A. N. et al. Genetic architecture of aluminum tolerance in rice (Oryza sativa) determined through genome-wide association analysis and QTL mapping. *PLoS Genet.*, 7, e1002221, https://doi.org/10.1371/journal.pgen.1002221 (2011).
4. Zhao, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications*, 13, 467, https://doi.org/10.1038/s41467-017-00206-9 (2017).
5. Huang, X. et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genet.*, 44, 32–39, https://doi.org/10.1038/ng1018 (2012a).
6. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 490, 497–501, https://doi.org/10.1038/nature11332 (2012b).
7. McCouch, S. R. et al. Open access resources for genome-wide association mapping in rice. *Nature Communications*, 7, 10532, https://doi.org/10.1038/ncomms10532 (2016).
8. Crowell, S. et al. Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. *Nature Communications*, 7, 10527, https://doi.org/10.1038/ncomms10527 (2016).
9. Mao, H. et al. Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. *Proc. Natl. Acad. Sci. USA*, 107, 19579–19584, https://doi.org/10.1073/pnas.1014191107 (2010).
10. Takano-Kai, N. et al. Evolutionary history of GS3, a gene conferring grain length in rice. *Genetics*, 182, 1323–1334, https://doi.org/10.1534/genetics.109.103002 (2009).
11. Weng, J. et al. Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res.* **18**, 1199–1209 (2008).
12. Si, L. et al. OsSPL13 controls grain size in cultivated rice. *Nature Genetics* **48**, 447–456, https://doi.org/10.1038/ng.3518 (2016).
13. Khush, G. S. Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35**, 25–34 (1997).
14. Liakat Ali, M. et al. A Rice Diversity Panel Evaluated for Genetic and Agro-Morphological Diversity between Subpopulations and its Geographic Distribution. *Crop Sci.* **51**, 2021, https://doi.org/10.2135/cropsci2010.11.0641 (2011).
15. Anmiraju, J. S. et al. Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *The Plant Cell* **20**, 3191–3209 (2008).
16. Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genet.* **42**, 961–967, https://doi.org.org/10.1038/ng.695 (2010).
17. Yano, K. et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genet.* **48**, 927–934, https://doi.org/10.1038/ng.3596 (2016).
18. Wang, X. et al. New Candidate Genes Affecting Rice Grain Appearance and Milling Quality Detected by Genome-Wide and Gene-Based Association Analyses. *Frontiers in Plant Science* **7**, 1998, https://doi.org/10.3389/fpls.2016.01998 (2016).
19. Ge, X., Xing, Y. Z., Xu, C. & He, Y. QTL analysis of cooked rice grain elongation, volume expansion, and water absorption using a recombinant inbred population. *Plant Breed.* **124**, 121–126 (2005).
20. Li, Y. et al. QTL analysis for cooking traits of super rice with a high-density SNP genetic map and fine mapping of a novel boiled grain length locus. *Plant Breed.* **134**, 535–541 (2015).
21. Amarawathi, Y. et al. Mapping of quantitative trait loci for basmati quality traits in rice (*Oryza sativa* L.). *Mol. Breed.* **21**, 49–65 (2008).
22. The 3000 Rice Genomes Project. The 3,000 rice genomes project.
23. Tamba, C. L., Ni, Y.-L. & Zhang, Y.-M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS computational biology* **13**, e1005357 (2017).
24. Wang, S. B. et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific reports* **6**, 19444, https://doi.org/10.1038/srep19444 (2016).
25. Litt, A. et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings in bioinformatics*, online, https://doi.org/10.1093/bib/bbw145 (2017).
26. Fan, C. et al. 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, https://doi.org/10.1007/s00122-006-0218-1 (2006).
27. Shomura, A. et al. Deletion in a gene associated with grain size increased yields during rice domestication. *Nature Genet.* **40**, 1023–1028 (2008).
28. Molina, J. et al. Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences, USA* **108**, 8351–8356 (2011).
29. Anaclet, R. et al. Prospects of breeding high-quality rice using post-genomic tools. *Theor. Appl. Genet.* **128**, 1449–1466 (2015).
30. Juliano, B. O. & Villareal, C. Grain quality evaluation of world rices. (Int. Rice Res. Inst., 1993).
31. Huang, R. et al. Genetic bases of rice grain shape: So many genes, so little known. *Trends Plant Sci.* **18**, 218–226, https://doi.org/10.1016/j.tplants.2012.11.001 (2013).
32. Huang, X., Lu, T. & Han, B. Resequencing rice genomes: an emerging new era of rice genomics. *Trends Genet.* **29**, 225–232, https://doi.org/10.1016/j.tig.2012.12.001 (2013).
33. Song, X.-J., Huang, W., Shi, M., Zhu, M.-Z. & Lin, H.-X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nature Genet.* **39**, 623–630 (2007).
34. Li, N. & Li, Y. Ubiquitin-mediated control of seed size in plants. *Front Plant Sci.* **5**, 332, https://doi.org/10.3389/fpls.2014.00332 (2014).
35. Ritt, C., Grimm, R., Fernández, S., Alonso, J. C. & Graser, K. D. Basic and acidic regions flanking the HMG domain of maize HMGa modulate the interactions with DNA and the self-association of the protein. *Biochemistry* **37**, 2673–2681 (1998).
36. Launhardt, D., Merkle, T., Houben, A., Schulz, A. & Grasser, K. D. Arabidopsis chromatin-associated HMGa and HMGb use different nuclear targeting signals and display highly dynamic localization within the nucleus. *The Plant Cell* **18**, 2904–2918, https://doi.org/10.1105/tpc.106.047274 (2006).
37. Kim, H. B. et al. Comprehensive analysis of AHL homologous genes encoding AT-hook motif nuclear localized protein in rice. *BMB Reports* **44**, 680–685, https://doi.org/10.5438/BMBRep.2011.44.10.680 (2011).
38. Antosch, M., Mortensen, S. A. & Grasser, K. D. Plant proteins containing high mobility group box DNA-binding domains modulate different nuclear processes. *Plant Physiol.* **159**, 875–883, https://doi.org/10.1104/pp.14.01199 (2014).
39. Franco-Zorrilla, J. M. et al. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences, USA* **111**, 2367–2372, https://doi.org/10.1073/pnas.1316278114 (2014).
40. Bandillo, N. et al. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice* **6**, 11 (2013).
41. Mathew, I. E., Das, S., Mahito, A. & Agarwal, P. Three Rice NAC Transcription Factors Heteromerize and Are Associated with Seed Size. *Frontiers in Plant Science* **7**, 1638, https://doi.org/10.3389/fpls.2016.01638 (2016).
42. Fang, N. et al. SMALL GRAIN 11 Controls Grain Size, Grain Number and Grain Yield in Rice. *The Plant Cell* **33**, 2904–2918, https://doi.org/10.1093/tig/tsw188 (2017).
43. Huang, X., Lu, G., Zhao, Q., Liu, X. & Han, B. Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant Physiol.* **148**, 25–40 (2008).
44. Yu, J. et al. Thames, R. S. et al. OsSPL13 controls grain size in cultivated rice. *Nature Genet.* **48**, 447–456, https://doi.org/10.1038/ng.3518 (2016).
45. Fang, N. et al. SMALL GRAIN 11 Controls Grain Size, Grain Number and Grain Yield in Rice. *Cell Res.* **26**, 1199–1209, https://doi.org/10.10115/cellularres.2008.307 (2008).
46. Yu, P. et al. Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* **12**, 372 (2011).
47. Li, M.-M. et al. Identification of quantitative trait loci for grain traits in japonica rice. *Agricultural Sciences in China* **9**, 929–936 (2010).
48. Schatz, M. C. et al. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* **15**, 506 (2014).
49. Mahesh, H. et al. Indica rice genome assembly, annotation and mining of blast disease resistance genes. *BMC Genomics* **17**, 242 (2016).
Acknowledgements
The authors thank Dr. Hei Leung, International Rice Research Institute, Philippines for insightful comments and helpful discussions. We acknowledge the Advanced Science and Technology Institute of Department of Science and Technology, Republic of the Philippines for hosting our biocomputational server, and providing additional storage space used during the genomic analysis. We also acknowledge the technical help of Ms. Lilia Molina and Grain Quality and Nutrition Service Lab team in measuring raw and cooked grain dimensions. We thank Ms. Sabhiha Parween and Dr. Vito Butardo for gene expression analysis of selected candidate gene from microarray experiment. We also thankfully acknowledge the Mr. Roven Fuentes for providing the Indel data for conducting GWAS. This work has been supported under the CGIAR thematic area Global Rice Science Partnership (GRISP), RICE CRP, Stress-Tolerant Rice for Africa and South Asia (STRASA) Phase III for BMGF funding.

Author Contributions
N.S. conceived the project and designed the research. N.S. and A.G. supervised the PhD work of G.M. G.M. performed synteny analysis. R.A. and G.M. conducted GWAS analysis and haplotype mining. S.B. interpreted the GWAS results. N.A. provided data used in indel analysis and genotyping calls based on SNP data. S.B., R.A. and N.S. wrote the manuscript with contributions from co-authors.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-12778-6.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017