Emergence of a Novel Chimeric Gene Underlying Grain Number in Rice

Hao Chen,[1] Yanyan Tang,[1] Jianfeng Liu, Lubin Tan, Jiahuan Jiang, Mumu Wang, Zuofeng Zhu, Xianyou Sun, and Chuanqing Sun[2]
State Key Laboratory of Plant Physiology and Biochemistry, National Center for Evaluation of Agricultural Wild Plants (Rice), Department of Plant Genetics and Breeding, China Agricultural University, Beijing 100193, China

ABSTRACT Grain number is an important factor in determining grain production of rice (*Oryza sativa* L.). The molecular genetic basis for grain number is complex. Discovering new genes involved in regulating rice grain number increases our knowledge regarding its molecular mechanisms and aids breeding programs. Here, we identified *GRAINS NUMBER 2* (*GN2*), a novel gene that is responsible for rice grain number, from "Yuanjiang" common wild rice (*O. rufipogon* Griff.). Transgenic plants overexpressing *GN2* showed less grain number, reduced plant height, and later heading date than control plants. Interestingly, *GN2* arose through the insertion of a 1094-bp sequence from *LOC_Os02g45150* into the third exon of *LOC_Os02g56630*, and the inserted sequence recruited its nearby sequence to generate the chimeric *GN2*. The gene structure and expression pattern of *GN2* were distinct from those of *LOC_Os02g45150* and *LOC_Os02g56630*. Sequence analysis showed that *GN2* may be generated in the natural population of Yuanjiang common wild rice. In this study, we identified a novel functional chimeric gene and also provided information regarding the molecular mechanisms regulating rice grain number.

KEYWORDS *GN2*; grain number; chimeric gene; gene emergence; rice

As a widely consumed staple food in the world, rice (*Oryza sativa* L.) is an important calorie source for humans. With the daily increase in the world's population and the decrease in potential available farmlands, effectively increasing rice grain production has become a growing challenge (Khush 1999; Zhang 2007). Rice grain production is thought to be determined by three major factors: panicle per plant, grain weight and grain number (Xing and Zhang 2010). Compared with the other two major factors, rice grain number exhibits a wider phenotypic variation, and is regarded as a main selection and improvement target (Yamagishi *et al.* 2002). Consequently, currently cultivated rice lines display dramatically increased grain number compared with its wild progenitor, the common wild rice (*O. rufipogon* Griff.) (Sun *et al.* 2001; Kovach *et al.* 2007).

Rice grain number is a complex agronomic trait impacted by many factors, including panicle architecture, and the initiation and outgrowth of branches and spikelets (Xing and Zhang 2010; Wang and Li 2011). Recently, a few quantitative trait loci (QTL) and genes regulating rice grain number have been identified. Among which *DEP1*, *DEP2*, *LP*, *SP1*, *PAP2*, and *sped1-D* alter panicle architecture (Huang *et al.* 2009; Li *et al.* 2009, 2010, 2011; Zhou *et al.* 2009; Kobayashi *et al.* 2010; Jiang *et al.* 2014), whereas *Gn1a*, *LAX1*, *SPA1*, *TAW1*, *DST*, and *FZP* control the initiation and outgrowth of branches and spikelets (M. Komatsu *et al.* 2003; K. Komatsu *et al.* 2003; Ashikari *et al.* 2005; Li *et al.* 2013; Yoshida *et al.* 2013). In addition, some genes, such as *Ghd7*, *Ghd8*, *PROG1*, *IPA1*, *FUWA*, *PAY1*, and *An1*, underlying rice grain number show pleiotropic effects in many significant agronomic- or domestication-related traits, including plant height, heading date, plant architecture, awn habit, and grain size (Jin *et al.* 2008; Tan *et al.* 2008; Xue *et al.* 2008; Jiao *et al.* 2010; Miura *et al.* 2010; Yan *et al.* 2011; Luo *et al.* 2013; Chen *et al.* 2015; Zhao *et al.* 2015).

Chimeric genes originate from multiple parental loci, and they have been reported in a variety of organisms (Long *et al.* 2003). Chimeric genes have the potential to evolve novel functions different from those of the parental loci, and provide the opportunity to study gene function and evolution

Copyright © 2017 by the Genetics Society of America
doi: 10.1534/genetics.116.188201
Manuscript received March 2, 2016; accepted for publication December 8, 2016; published Early Online December 16, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.188201/-/DC1.
[1]These authors contributed equally to this work.
[2]Corresponding author: State Key Laboratory of Plant Physiology and Biochemistry, China Agricultural University, Beijing 100193, China. E-mail: suncq@cau.edu.cn

(Long *et al.* 2003). Some chimeric genes generate novel functions and influence phenotypes. For example, *sphinx* determines male courtship in *Drosophila* (Wang *et al.* 2002), *CYPA-TRIM5* affects HIV1-resistance in primates (Sayah *et al.* 2004), *CHANNEL11/12* regulates multiple pathogen-resistance responses in *Arabidopsis* (Yoshioka *et al.* 2006), *BS1* takes part in the normal reproductive development of maize (Jin and Bennetzen 1989; Elrouby and Bureau 2010), and chimeric *ATP6* and *Rfo* lead to male sterility in rice and radish, respectively (Kadowaki *et al.* 1990; Brown *et al.* 2003; Z. Wang *et al.* 2006). In the rice genome, chimeric genes have also been widely detected (W. Wang *et al.* 2006). Systematic analysis demonstrated that chimeric genes account for a high percentage of the total newly evolved genes (W. Wang *et al.* 2006; Fan *et al.* 2008; Zhang *et al.* 2013). However, the phenotypic effects of most rice chimeric genes have not been well-investigated.

In this study, we identified a novel chimeric gene, *GRAINS NUMBER 2* (*GN2*), which is responsible for rice grain number, as well as plant height and heading date, on the long arm of chromosome 2 in Yuanjiang common wild rice. We found that a 1094-bp sequence from *LOC_Os02g45150* was inserted into the third exon of *LOC_Os02g56630*, and that the inserted sequence recruited its nearby sequence to generate the chimeric *GN2*. *GN2* was generated in the natural population of Yuanjiang common wild rice. The discovery of *GN2* could be useful in elucidating the molecular mechanisms underlying rice grain number and in uncovering information regarding gene emergence.

Materials and Methods

Plant materials

"YIL19" is a rice introgression line derived from a set of 106 introgression lines developed previously (Tan *et al.* 2007). This set of introgression lines was derived from a cross between an accession of Chinese common wild rice (Yuanjiang, *O. rufipogon* Griff.) and a high-yield *indica* cultivar (*O. sativa* L.) "Teqing" followed by four generations of selfing and three generations of backcrossing. Yuanjiang common wild rice was collected from Yuanjiang Country, Yunnan Province, China.

Map-based cloning

An F_2 population containing 190 individual plants derived from the cross between YIL19 and Teqing was constructed for QTL analysis. QTL analysis was carried out by interval analysis using Map Manager QTXb17 (Manly *et al.* 2001). An F_2 segregating population containing 4512 plants (including the previous 190 F_2 plants) was used for the genetic analysis and fine-mapping. The F_3 families, derived from the F_2 recombinant individuals, were planted to confirm the phenotypes of the F_2 recombinant individuals. Some molecular markers used for map-based cloning were published markers (McCouch *et al.* 2002). The newly developed molecular markers used in this study are listed in Supplemental Material, Table S4.

Gene prediction and annotation

The annotation of putative genes in the fine-mapping region were obtained from the Rice Genome Annotation Project (RGAP, MSU V7; http://rice.plantbiology.msu.edu/downloads.shtml). We also annotated genes according to the sequence information of YIL19 and Teqing from the Rice Genome Automated Annotation System (http://RiceGAAS.dna.affrc.go.jp/).

Detection of the possible transcripts by strand-specific reverse transcription-polymerase chain reaction (RT-PCR) analysis

To detect the possible transcript generated by the 1094-bp DNA fragment insertion in the locus of *LOC_Os02g56630*, 22 pairs of specific primers covering a 3-kb genomic region surrounding the insertion site were designed to analyze the complementary DNA (cDNA) of YIL19 and Teqing. The RT-PCR products were confirmed by gel electrophoresis and sequence analyses.

Real-time quantitative PCR (RT-qPCR) and rapid amplification of cDNA ends (RACE) analysis

Total RNA was extracted from various samples, including 7-day-old seedlings, roots, leaves, leaf sheathes, tiller bases, shoot apical meristems, young panicles shorter than 1 cm, young panicles longer than 1 cm and shorter than 5 cm, young panicles longer than 5 cm, mature panicles, and spikelet hulls, using TRIzol reagent (Invitrogen, Carlsbad, CA, http://www.lifetechnologies.com) and was purified using an RNeasy Micro Kit (QIAGEN, Valencia, CA, https://www.qiagen.com). For RT-qPCR analyses, first-strand cDNA was synthesized using an oligo(dT)18 primer (TaKaRa, http://www.takara.com) and SuperScript III Reverse Transcriptase (Invitrogen) from 2 μg of total RNA. The expression levels of *GN2* and other genes were analyzed using a CFX96 Real Time System (Bio-Rad, Hercules, CA, http://www.bio-rad.com), and the rice *Ubiquitin* gene (*LOC_Os03g13170*) served as the control. Each set of experiments was repeated three times. 5'- and 3'-RACE were carried out using a 5'-Full RACE Kit with tobacco acid pyrophosphatase (TAP) and a 3'-Full RACE Core Set with PrimeScript RTase (TaKaRa) according to the manufacturer's instructions.

Complementation test

The constructs pUbi::*GN2.1* and pUbi::*GN2.2* harbored the full-length sequence of *GN2.1* and *GN2.2* cDNA, respectively, under the control of a maize *Ubiquitin* promoter. These constructs were introduced into *Agrobacterium tumefaciens* strain EHA105 and were subsequently transferred into Teqing. Two independent lines per construct were used for further phenotypic evaluations.

Detection of the presence of GN2 in rice germplasm

We detected the presence of *GN2* by two methods. First, the nucleotide BLAST search was performed using the genomic sequence and full-length cDNA sequence of *GN2* against the

corresponding databases. Second, a pair of diagnostic PCR primers flanking the 1094-bp DNA fragment was designed to detect the presence of *GN2* in 236 rice germplasm accessions (Table S7). If *GN2* is present, then the expected PCR amplification band size between the diagnostic primers is ~2 kb, while if *GN2* is absent, then the expected PCR amplification band size is ~1 kb.

Protoplast transfection assays for regulatory sequence analysis

The sequence P1 contains the ancestral sequence of *LOC_Os02g56630*, including a ~2-kb genomic sequence upstream of *GN2*, and P2 contains the upstream sequence of *GN2* from the insert sequence of *LOC_Os02g45150*. All of the sequences were amplified by PCR from YIL19 genomic DNA and then inserted in the frame of pGreenII 0800-LUC (Hellens *et al.* 2005). Protoplasts were prepared from 7-day-old seedlings of Teqing. After protoplast preparation and transfection, firefly luciferase (LUC) and *Renilla* luciferase (REN) activities were measured using the Dual-Luciferase Reporter Assay System (Promega, Madison, WI, http://www.promega.com). Relative LUC activity was calculated by normalizing against the REN activity following the manufacturer's instructions. In the experiments, 10 μg of each of the constructs was used, and all of the experiments were repeated three times with similar results.

Data availability

The authors state that all data necessary for confirming the conclusions presented in this article are represented fully within the article. Plant materials in this study are available upon request.

Results

YIL19 shows less grain number and pleiotropic phenotypes

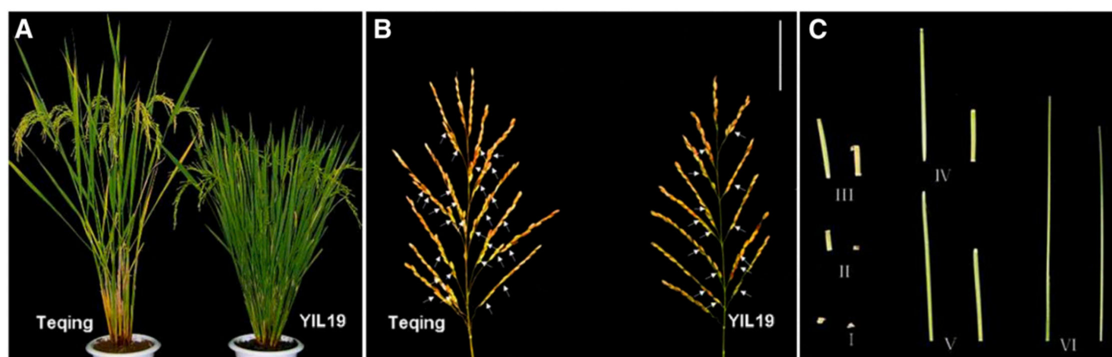
An introgression line, YIL19, which originated from a set of introgression lines constructed using Yuanjiang common wild rice as the donor parent and the *indica* variety Teqing as the recurrent parent, exhibited less grain number compared with its recurrent parent (Figure 1, A–C). The investigation of panicle traits showed that the grain number per panicle for YIL19 and Teqing was 85.7 and 155.3, respectively (Figure 1D). Significant differences ($P < 0.01$) were detected in grain number per panicle, panicle length, mean length of primary branches, number of secondary branches, and grains on secondary branches between YIL19 and Teqing, while no significant difference was present in the number of primary branches (Figure 1, E–I and Figure S1). An evaluation of other important agronomic traits between YIL19 and Teqing indicated that there were significant differences ($P < 0.01$) in plant height, days to heading, panicle number per plant, and grain yield per plant; however, no significant differences were detected in grain weight or spikelet fertility (Figure 1, J and K and Table S1). In addition, all of the internodes of the stems were shortened in the YIL19 plants compared with the Teqing plants (Figure 1C).

Detecting and fine-mapping GN2

YIL19 contained four *O. rufipogon* chromosomal segments, located on chromosomes 1, 2, 6, and 10 (Figure 2A). The grain number of the F_1 hybrid derived from the cross between YIL19 and Teqing was similar to that of YIL19 (Figure 2B). An F_2 population containing 190 individuals was genotyped using 14 simple sequence repeat markers distributed on four *O. rufipogon* introgression regions and evaluated for grain number on the main panicle. QTL analysis revealed that there was a QTL between markers RM5300 and RM6312 on the long arm of chromosome 2 with a LOD score of 12.24 that explained 26% of the phenotypic variance (Table S2 and Table S5). The *O. rufipogon*-derived allele contributed a decreasing effect on grain number. We referred to this locus as *GN2*. No QTL underlying rice grain number was detected in the other introgression regions, suggesting that *GN2* was responsible for the less grain number phenotype of YIL19 (Figure 2C). Using 4512 F_2 individuals, the location of *GN2* was narrowed down to an interval of 47 kb between RM3535 and the newly developed molecular marker R4 (Figure 2C, Figure S2, and Table S6). There were six predicted genes in this fine-mapping region based on the "Nipponbare" genome (Figure 2, D and E and Table S3).

The newly detected transcript in YIL19 is the major GN2 candidate

A sequence analysis of the six predicted genes revealed that a 1094-bp DNA fragment insertion was present at position +2275 on exon 3 of *LOC_Os02g56630* in the YIL19 genome, when compared with the sequence of Teqing (Figure 2F), while no such large variation occurred in the other five genes. Furthermore, this 1094-bp DNA fragment insertion cosegregated with the less grain number phenotype. The annotation analysis also suggested that the 1094-bp DNA fragment insertion in YIL19 resulted in the premature termination of *LOC_Os02g56630* messenger RNA (mRNA) translation. Therefore, *LOC_Os02g56630*, with three exons and two introns, which encodes an OsWAK (Wall-associated kinase) receptor-like protein kinase of 765 amino acids, appears to be a *GN2* candidate. However, the expression of *LOC_Os02g56630* cannot be detected in all of the tested tissues collected, which span the whole rice life cycle, in either YIL19 or Teqing. Therefore, we speculated that other unpredicted transcripts in this region may influence the rice grain number. To detect the possible transcripts, we designed 22 pairs of primers covering a ~3-kb genomic region surrounding the insert site to conduct the strand-specific RT-PCR analysis. A ~200-bp transcript spanning the inserted sequence and the third exon of *LOC_Os02g56630* was detected in YIL19 (Figure S3), while no transcript was detected from the corresponding locus of Teqing. To determine the full length of the transcript, 5′- and 3′-RACE was conducted. We identified two unique transcripts, *GN2.1* and *GN2.2*. Sequence analysis revealed that *GN2.1* was 497 bp in length, while *GN2.2* was 494 bp, with only a 3-bp deletion difference from *GN2.1*



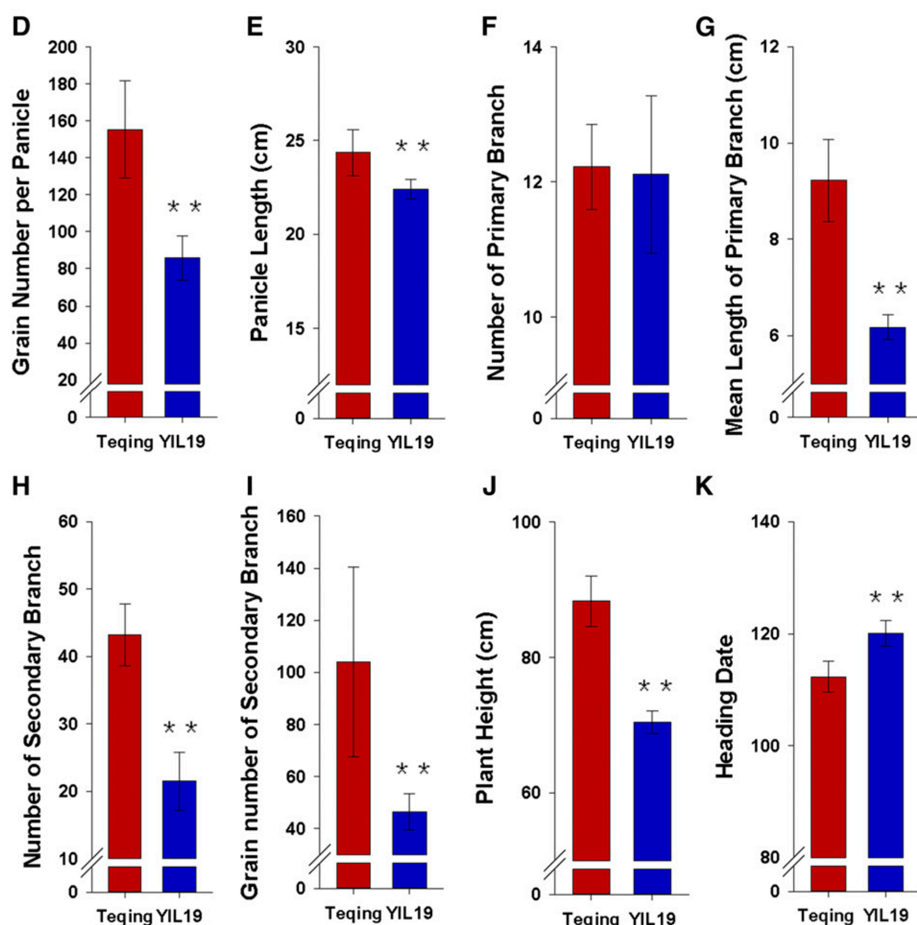


Figure 1 Phenotypes of YIL19 and Teqing. (A) Phenotypes of YIL19 and Teqing plants at the maturity stage. (B) The panicle structures of YIL19 and Teqing. The white arrows represents the place where a secondary branch emergences. Scale bar = 5 cm. (C) Stem structures of YIL19 and Teqing. On the left is the length of each internode in Teqing, and on the right is YIL19. (D–K) Comparison of grain number per panicle, panicle length, number of main panicles, mean length of primary branches, secondary branch number, grain number of secondary branches, plant height, and heading date between YIL19 and Teqing. All of the data above are given as mean ± SE (n = 20). (D–K) ** $P < 0.01$, determined by the Student's t-test.

(Figure 2G and Figure S4), suggesting that *GN2.1* and *GN2.2* were transcript variants of the same gene, which has two exons and one intron (Figure 2H). Further analyses demonstrated that the 3-bp deletion in *GN2.2* occurs at position +22 of the cDNA sequence, and is caused by an alternative recognition of the RNA splice junction "AG" at the 5′-end of the second exon (Figure 2I). In addition, *GN2.1* and *GN2.2* showed a chimeric structure composed of 95 and 92 bp, respectively, from the 1094-bp inserted sequence and 402 bp from *LOC_Os02g56630* (Figure 2H and Figure S5).

Complementation test of GN2

To verify whether the novel transcripts resulted in the less grain number phenotype of YIL19, we performed a transgenic

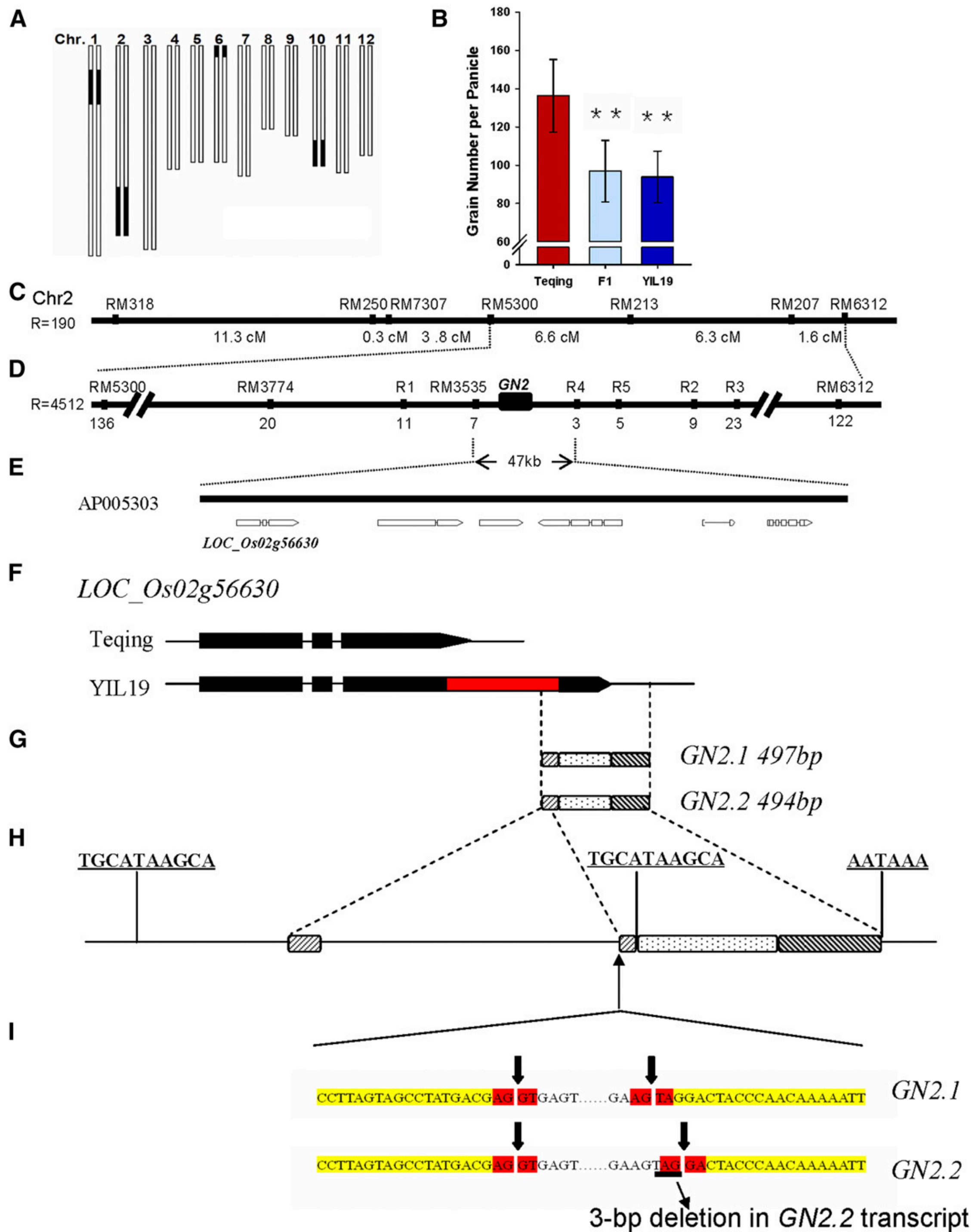


Figure 2 Fine-mapping and gene candidates for *GN2*. (A) The introgression condition of YIL19. The boxes represent rice chromosomes: solid boxes represent the chromosome segment from Yuanjiang common wild rice and the open boxes represent the segments from Teqing. (B) Grain number of YIL19 and Teqing, and their heterosis. (C) Primary mapping of *GN2* between markers RM318 and RM6312. *R* represents the number of individuals in a population. (D) Fine-mapping of *GN2* harboring a 47-kb region of AP005303. (E) The putative candidate genes for *GN2* based on the annotation of Nipponbare. The open boxes represent the exons of genes based on the annotated information. (F) The gene structures of *LOC_Os02g56630* in YIL19 and Teqing. The black box indicates the exon of *LOC_Os02g56630*; the black lines indicate introns and intergenic regions, while the red box indicates the inserted DNA segment of *LOC_Os02g56630* in YIL19. (G) The mRNA splices of new transcripts. The boxes indicate the exons of the new transcripts, while different types of stripes and dots in the boxes represent the sequences having different origins. (H) The gene structure of the new transcripts. The boxes indicate the exons of new transcripts, while black lines among the boxes are the introns and intergenic sequences. The direct repeat sequences are marked by TGCATAAGCA, while the polyadenylation sites are marked by AATAAA. (I) The alternative splicing style of the new transcripts. The sequence in yellow is the exon of the new transcripts; the sequence in red represents the splicing site. The black arrows are pointing to the splicing site. The 3-bp deletion sequence in *GN2.2* is underlined with a black line. Chr, chromosome; mRNA, messenger RNA.

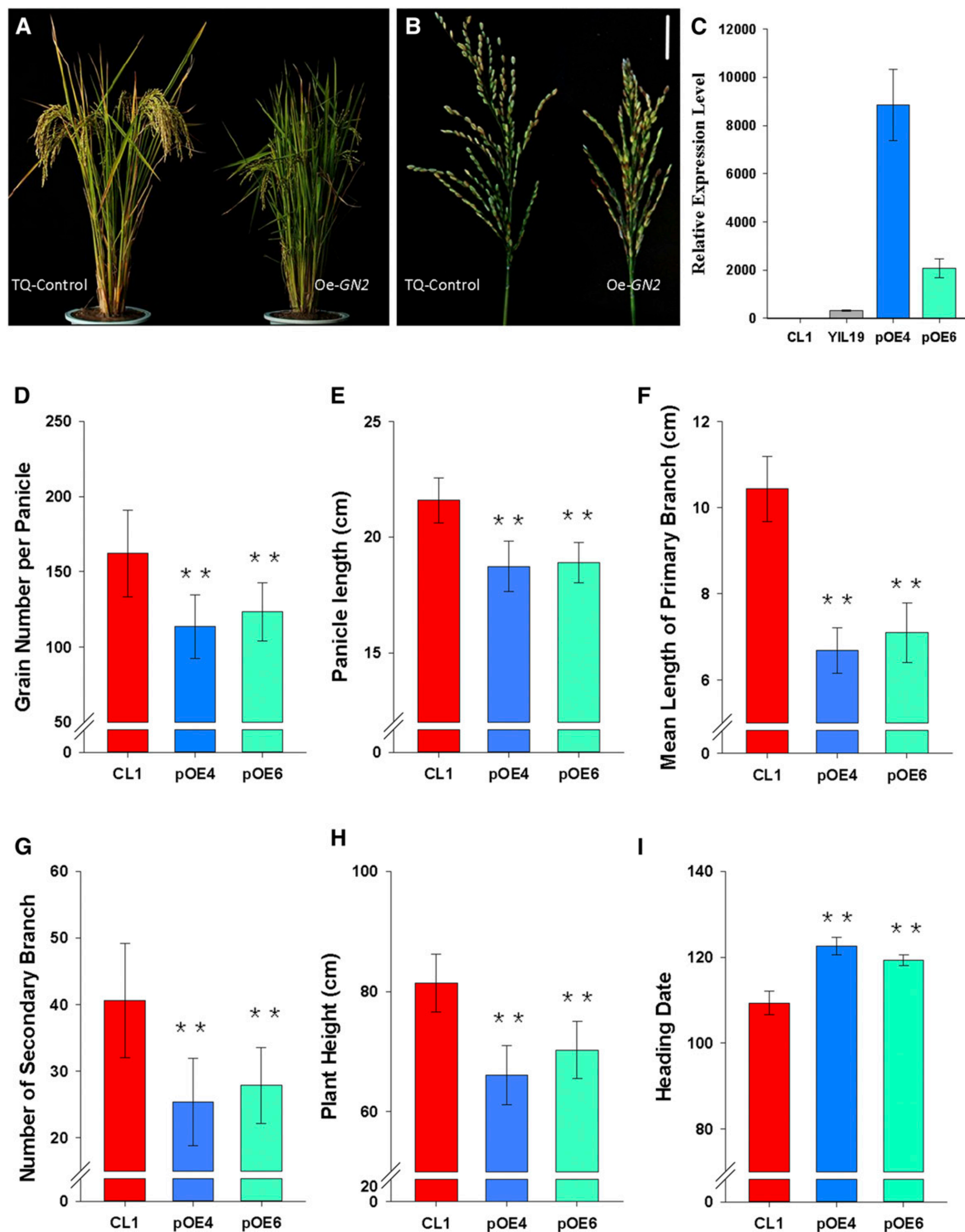


Figure 3 Complementation test of *GN2*. (A) Phenotype comparison between the *GN2*-overexpressing transgenic plants and control plants transformed with an empty plasmid. (B) Panicle comparison between the *GN2*-overexpressing transgenic plants and control plants transformed with an empty plasmid. Oe-*GN2* here represents the *GN2.1*-overexpressing transgenic line. Scale bar = 5 cm. (C) The relative expression levels of *GN2* in different *GN2*-overexpressing transgenic plants. CL1 represents the Teqing control line with an empty plasmid used for analysis. *pOE4* and *pOE6* represent *GN2.1*-overexpressing transgenic lines used for analysis. (D–I) Comparisons of grain number per panicle, panicle length, average length of primary branches, number of secondary branches, plant height, and heading date among *GN2*-overexpressing transgenic lines and their control lines. All of the data above are presented as mean ± SE (n = 20). ** $P < 0.01$, determined by the Student's t-test.

analysis. A construct overexpressing the novel transcript *GN2.1*, driven by a maize *Ubiquitin* promoter, was introduced into Teqing plants. Twelve overexpressing transgenic lines (designated as *GN2.1*-Teqing) were obtained (Figure 3, A and B). A RT-qPCR analysis showed that these transgenic lines overexpressed *GN2.1* transcripts compared with the

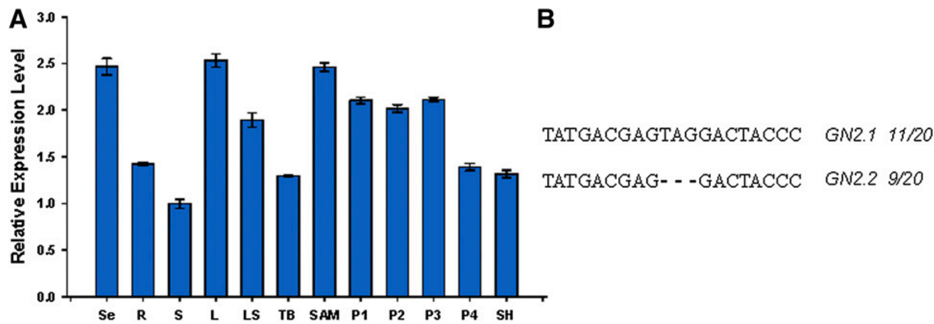


Figure 4 The spatial expression pattern of *GN2*. (A) The spatial expression pattern of *GN2*. Values are means ± SD (B) The occurrence frequencies of different *GN2* transcripts. The dot indicates the 3-bp deletion in *GN2.2*. L, leaf; LS, leaf sheath; P1, young panicle shorter than 1 cm; P2, young panicle longer than 1 cm and shorter than 5 cm; P3, young panicle longer than 5 cm; P4 mature panicle; R, root; S, stem; SAM, shoot apical meristem; Se, seedling; SH, spikelet hull; TB, tiller base.

control lines (Figure 3C). All of the *GN2.1*-Teqing transgenic lines displayed less grain number than control lines. Significant differences ($P < 0.01$) were detected in grains per panicle, panicle length, average length of primary branches, and number of secondary branches, compared with the control lines (Figure 3, A, B, and D–G). Additionally, *GN2.1*-Teqing transgenic lines showed reduced plant height and delayed heading date compared with their control lines (Figure 3, H and I).

We also developed a construct overexpressing the *GN2.2* transcript and introduced it into Teqing (designated as *GN2.2*-Teqing). Ten independent transgenic lines were obtained. Similar phenotypes were observed in the transgenic plants containing these two different constructs (Figure S6). Thus, the novel gene, with two alternative transcripts, corresponded to *GN2*, and influenced the grain number and other plant agronomic-related traits in rice.

Expression pattern of GN2

To detect the expression pattern of *GN2*, we surveyed its expression profiles in YIL19 using the RNA from tissues spanning the whole life cycle, including 7-day-old seedlings, roots, leaves, leaf sheathes, tiller bases, shoot apical meristems, young panicles from different stages, mature panicles, and spikelet hulls. *GN2* was constitutively expressed in all of the surveyed tissues. However, the relative expression level of *GN2* appeared to be low in the tested tissues (Figure 4A).

To determine the occurrence ratio of the two independent transcripts, we randomly sequenced 20 *GN2* clones. Eleven clones expressed *GN2.1*, and the remaining nine clones expressed *GN2.2*, suggesting that *GN2.1* and *GN2.2* have similar expression levels (Figure 4B).

GN2 may influence rice grain number using a distinct pathway from known grain number regulators

Many genes associated with grain number have been identified in rice (Xing and Zhang 2010). To understand whether these genes are regulated indirectly by *GN2*, we carried out a RT-qPCR analysis to investigate their expression profiles. The expression levels of most of the tested genes did not show significant differences between YIL19 and Teqing in young panicles (Figure S7). Interestingly, the expression level of *TAW1*, and its related genes *OsMADS22* and *OsMADS55*, showed a higher expression level in young panicles of YIL19 compared with in Teqing (Figure S7). However, a previous study revealed that the higher expression level of *TAW1* led to an increase in the number of secondary branches (Yoshida *et al.* 2013). Thus, *GN2* may influence rice grain number through a novel pathway distinct from those genes reported previously.

Emergence of the chimeric GN2 gene

To trace the emergence of *GN2*, we conducted a BLAST search against the *O. sativa* L. ssp. *japonica* genome set and found that 1074 bp of the 1094-bp inserted DNA fragment had a very high DNA sequence identity (> 99%) with the sequence of *LOC_Os02g45150* in Nipponbare. The remaining 20 bp is the direct repetitive sequence "TGCATAAGCA" of unknown origin that flanks the 1074-bp sequence. Further analysis showed that the 1074-bp sequence belongs to the partial 5′-untranslated region sequence (or partial 5′-untranslated region and intron according to another annotation version) of *LOC_Os02g45150* on chromosome 2 (Figure 5, A and B). Interestingly, the 1074-bp sequence in *LOC_Os02g45150* described above was absent in the genome of C4, the accession of Yuanjiang common wild rice that is the parent of the introgression line used in this study, suggesting that the inserted sequence may be "cut" from the locus of *LOC_Os02g45150* and then "pasted" into the locus of *LOC_Os02g56630* in Yuanjiang common wild rice (Figure 5C and Figure S9).

Based on the *GN2* gene construct, we inferred that the 1094-bp inserted sequence from the locus of *LOC_Os02g45150* contributed the transcriptional start site, the first exon, the whole intron, and the partial second exon of *GN2*, and *LOC_Os02g56630* provided the rest of the second exon and the polyadenylation [poly(A)] site (Figure 5D). Therefore, we concluded that *GN2* is a chimeric gene produced by the fusion of the two sequences from different origins.

To analyze the possible transcriptional regulatory region of *GN2*, a transient expression assay was conducted. The inserted sequence from *LOC_Os02g45150* had a weak regulatory ability, while the upstream sequence of *GN2* from *LOC_Os02g56630* did not, indicating that *GN2* acquired its regulatory sequences from the 1094-bp inserted sequence (Figure S8).

A BLAST similarity search showed that *GN2* is unique in the rice genome. It also revealed that *GN2* has no homolog in any other organisms. We further analyzed the presence of *GN2* in 236 rice accessions representing diverse germplasms,

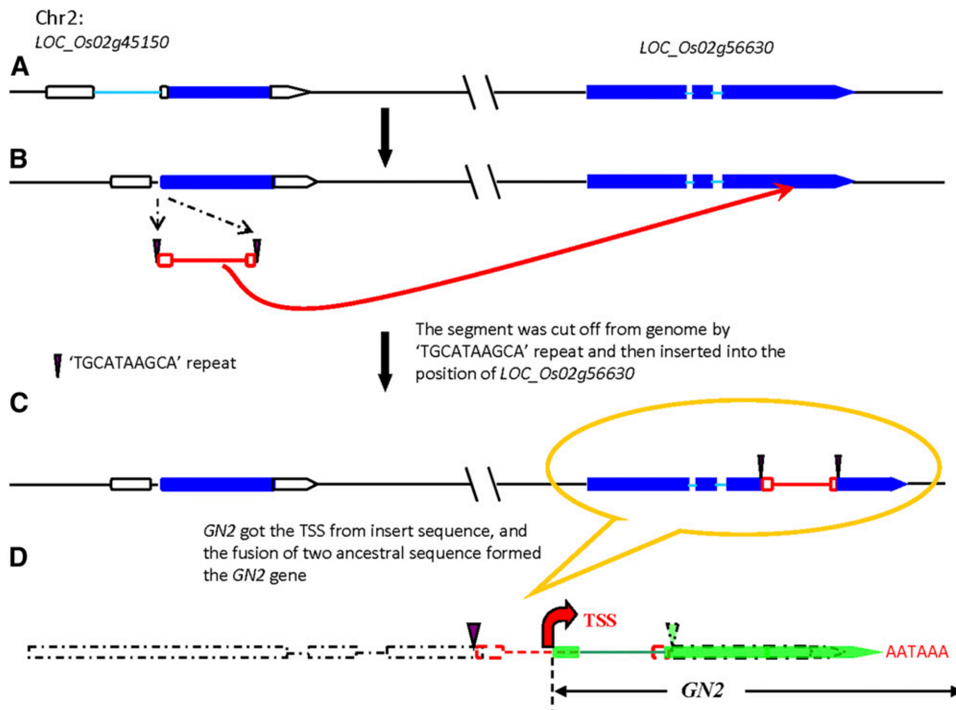


Figure 5 Emergence process of *GN2*. (A) Initial gene structures of *LOC_Os02g45150* and *LOC_Os02g56630*. The boxes indicate the exons of the genes; the dark blue box indicates the CDS region based on the annotation of Nipponbare, while the connecting line in dark blue represents the intron. The black connecting line and delimiter represent chromosome regions. (B) The movement of the DNA fragment from *LOC_Os02g45150* to the putative exon of *LOC_Os02g56630*. The box and connecting line in red represent the "sheared" DNA fragment in *LOC_Os02g45150*, and the triangles in purple represent the "TGCATAAGCA" repeats. (C) The gene structures of *LOC_Os02g45150* and *LOC_Os02g56630* after the insertion event. (D) The emergence of the *GN2* gene. The box and triangle in green represent the exons of *GN2*, and the connecting line in green represents the intron of *GN2*. The dotted line in black represents the ancestral gene structure of *LOC_Os02g56630*, and the dotted line in red represents the inserted genomic sequence. The red right angle arrow represents the TSS, and the AATAAA represents the polyadenylation site of *GN2*. CDS, coding sequence; Chr, chromosome; TSS, transcription start site.

including both wild and cultivated species (Table S7). None of the tested accessions exhibited the presence of *GN2*, suggesting that *GN2* may be newly generated in the genome of Yuanjiang common wild rice.

To detect the occurrence of *GN2* in the Yuanjiang common wild rice population, we analyzed the presence of *GN2* in 22 individuals from the local natural population of Yuanjiang common wild rice in Yunnan Province, China. We found that 12 individuals contained homozygous *GN2*, four individuals exhibited heterozygosity at the *GN2* locus, and six individuals did not contain *GN2*, suggesting that *GN2* has a high frequency in the local natural population of Yuanjiang common wild rice, although it is not completely fixed in this population.

Discussion

In this study, we identified a novel gene, *GN2*, underlying rice grain number, using an introgression line, YIL19. *GN2* exhibits pleiotropism, not only influencing grain number, but also affecting the plant height and heading date. Interestingly, *GN2* is a novel chimeric gene that emerged through the shuffling of a 1094-bp genomic fragment from *LOC_Os02g45150* to *LOC_Os02g56630* on the same chromosome. We also found that *GN2* is unique to the rice genome and newly emerged in the natural population of Yuanjiang common wild rice.

The insertion event produced new *GN2* transcripts. Sequence analysis revealed that the longest predicted open reading frame in the *GN2* transcript was 213-bp, encoding a 70-amino acid polypeptide. Therefore, we speculated that *GN2* may exert its function as a novel polypeptide.

The panicle branch plays an important role in determining rice grain number (Wang and Li 2011). Both the branch meristem's activities and the elongation of branches regulate rice grain number. Previous studies showed that a couple of genes, such as *LAX1*, *Gn1a*, and *TAW1*, influence rice grain number by regulating the number of branches (K. Komatsu *et al.* 2003; Ashikari *et al.* 2005; Yoshida *et al.* 2013). Several genes, such as *SP1*, control rice grain number by regulating the elongation of panicles, primary branches, and secondary branches (Li *et al.* 2010), while *sped1-D* influences the length of pedicels and secondary branches (Jiang *et al.* 2014). Unlike the genes described above, *GN2* has a specific effect on the elongation of the primary branches. The shortened length of the primary branches leads to a lower number of secondary branches in YIL19. This suggested that *GN2* regulates rice grain number in a manner distinct from those of known genes.

Chimeric genes are able to generate new functions that differ from those of parent sequences (Long *et al.* 2003). Although chimeric genes form a certain proportion of the rice genome, and many of them have expression profiles and undergo selection, their phenotypic effects are mostly unknown (W. Wang *et al.* 2006; Fan *et al.* 2008; Zhang *et al.* 2013). In this study, we verified that *GN2* is a novel chimeric gene that exhibits a distinctive gene structure and expression pattern from its parental genes, and we also found that *GN2* acquires novel functions and has a significant effect on morphological traits.

GN2 arose through the insertion of a 1094-bp sequence that may come from *LOC_Os02g45150*. Its emergence is consistent with the "cut and paste" DNA fragment shuffling model, having the direct repeat sequence "TGCATAAGCA" flanking the inserted region. RNA-mediated retroposition and DNA-mediated recombination are known to be two major mechanisms for chimeric gene emergence (Arguello *et al.* 2007; Kaessmann 2010). Most of the rice chimeric genes arose through one of the two mechanisms (Turcotte *et al.* 2001; Bennetzen 2005; W. Wang *et al.* 2006). Although we detected the flanking direct repeat, one of the hallmarks of retroposition, in *GN2*, we were unable to find other structural hallmarks of retroposition, such as the loss of intron(s) and the presence of poly(A) tracts. Furthermore, the repeat sequence "TGCATAAGCA" does not resemble any characterized transposon or retrotransposon, and this insertion event induced by the "TGCATAAGCA" repeat sequence has rarely been detected in the rice genome. Therefore, we speculated that the formation of the chimeric *GN2* may not have occurred through the mechanisms of RNA-mediated retroposition or DNA-mediated recombination, but through a different approach of chimeric gene origination.

Comparative genomic studies revealed that the emergence of novel genes in the rice genome occurred at a higher frequency and more rapid pace than other organisms, such as primates, and the process of gene emergence appears to be ongoing (Ge *et al.* 1999; Vaughan *et al.* 2003; W. Wang *et al.* 2006). Interestingly, *GN2* is only detected in the genome of Yuanjiang common wild rice. Furthermore, it exhibited a high frequency (up to 63%) in the natural population of Yuanjiang common wild rice. Hence, we speculated that *GN2* had newly emerged in this natural population. The *GN2* high frequency in the local natural population may be subject to positive selection, and thus *GN2* may be functional.

Many newly emerged genes, which have evolved novel functions, are involved in development, reproduction, and stress responses in various organisms (Tautz and Domazet-Loso 2011; Chen *et al.* 2013). These novel genes have the potential to be used in the improvement of important agronomic traits and broaden germplasm resources for breeding. Furthermore, understanding the mechanism of gene emergence will shed light on artificial synthesis of functional genes in the future. The identification of the newly emerged *GN2*, which obviously influences agronomic traits, provides more information to elucidate the molecular mechanisms underlying rice grain number and new insight into general phenotypic evolution.

Acknowledgments

We thank Dr. Qingwen Yang, the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences for providing the wild rice samples. This research was supported by the National Natural Science Foundation of China (grants 91535301 and 31171522) and the National Key R&D Program for Crop Breeding (2016YFD0100301).

Literature Cited

Arguello, J., C. Fan, W. Wang, and M. Long, 2007 Origination of chimeric genes through DNA-level recombination. Genome Dyn. 3: 131–146.

Ashikari, M., H. Sakakibara, S. Lin, T. Yamamoto, T. Takashi *et al.*, 2005 Cytokinin oxidase regulates rice grain production. Science 309: 741–745.

Bennetzen, J. L., 2005 Transposable elements, gene creation and genome rearrangement in flowering plants. Curr. Opin. Genet. Dev. 15: 621–627.

Brown, G. G., N. Formanová, H. Jin, R. Wargachuk, C. Dendy *et al.*, 2003 The radish *Rfo* restorer gene of *Ogura cytoplasmic* male sterility encodes a protein with multiple pentatricopeptide repeats. Plant J. 35: 262–272.

Chen, J., H. Gao, X. M. Zheng, M. Jin, J. F. Weng *et al.*, 2015 An evolutionarily conserved gene, *FUWA*, plays a role in determining panicle architecture, grain shape and grain weight in rice. Plant J. 83: 427–438.

Chen, S., B. H. Krinsky, and M. Long, 2013 New genes as drivers of phenotypic evolution. Nat. Rev. Genet. 14: 645–660.

Elrouby, N., and T. E. Bureau, 2010 *Bs1*, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. Plant Physiol. 153: 1413–1424.

Fan, C., Y. Zhang, Y. Yu, S. Rounsley, M. Long *et al.*, 2008 The subtelomere of *Oryza sativa* chromosome 3 short arm as a hot bed of new gene origination in rice. Mol. Plant 1: 839–850.

Ge, S., T. Sang, B. R. Lu, and D. Y. Hong, 1999 Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc. Natl. Acad. Sci. USA 96: 14400–14405.

Hellens, R. P., A. C. Allan, E. N. Friel, K. Bolitho, K. Grafton *et al.*, 2005 Transient expression vectors for functional genomics, quantification of promoter activity and RNA silencing in plants. Plant Methods 1: 13.

Huang, X., Q. Qian, Z. Liu, H. Sun, S. He *et al.*, 2009 Natural variation at the *DEP1* locus enhances grain yield in rice. Nat. Genet. 41: 494–497.

Jiang, G., Y. Xiang, J. Zhao, D. Yin, X. Zhao *et al.*, 2014 Regulation of inflorescence branch development in rice through a novel pathway involving the pentatricopeptide repeat protein *sped1-D*. Genetics 197: 1395–1407.

Jiao, Y., Y. Wang, D. Xue, J. Wang, M. Yan *et al.*, 2010 Regulation of *OsSPL14* by OsmiR156 defines ideal plant architecture in rice. Nat. Genet. 42: 541–544.

Jin, Y. K., and J. L. Bennetzen, 1989 Structure and coding properties of *Bs1*, a maize retrovirus-like transposon. Proc. Natl. Acad. Sci. USA 86: 6235–6239.

Jin, J., W. Huang, J. P. Gao, J. Yang, M. Shi *et al.*, 2008 Genetic control of rice plant architecture under domestication. Nat. Genet. 40: 1365–1369.

Kadowaki, K. I., T. Suzuki, and S. Kazama, 1990 A chimeric gene containing the 5′ portion of *atp6* is associated with cytoplasmic male-sterility of rice. Mol. Gen. Genet. 224: 10–16.

Kaessmann, H., 2010 Origins, evolution, and phenotypic impact of new genes. Genome Res. 20: 1313–1326.

Khush, G. S., 1999 Green revolution: preparing for the 21st century. Genome 42: 646–655.

Kobayashi, K., M. Maekawa, A. Miyao, H. Hirochika, and J. Kyozuka, 2010 *PANICLE PHYTOMER2 (PAP2)*, encoding a SEPALLATA subfamily MADS-box protein, positively controls spikelet meristem identity in rice. Plant Cell Physiol. 51: 47–57.

Komatsu, K., M. Maekawa, S. Ujiie, Y. Satake, I. Furutani *et al.*, 2003 *LAX* and *SPA*: major regulators of shoot branching in rice. Proc. Natl. Acad. Sci. USA 100: 11765–11770.

Komatsu, M., A. Chujo, Y. Nagato, K. Shimamoto, and J. Kyozuka, 2003 *FRIZZY PANICLE* is required to prevent the formation of

axillary meristems and to establish floral meristem identity in rice spikelets. Development 130: 3841–3850.
Kovach, M. J., M. T. Sweeney, and S. R. McCouch, 2007 New insights into the history of rice domestication. Trends Genet. 23: 578–587.
Li, F., W. Liu, J. Tang, J. Chen, H. Tong *et al.*, 2010 Rice *DENSE AND ERECT PANICLE 2* is essential for determining panicle outgrowth and elongation. Cell Res. 20: 838–849.
Li, M., D. Tang, K. Wang, X. Wu, L. Lu *et al.*, 2011 Mutations in the F-box gene *LARGER PANICLE* improve the panicle architecture and enhance the grain yield in rice. Plant Biotechnol. J. 9: 1002–1013.
Li, S., Q. Qian, Z. Fu, D. Zeng, X. Meng *et al.*, 2009 *Short panicle1* encodes a putative PTR family transporter and determines rice panicle size. Plant J. 58: 592–605.
Li, S., B. Zhao, D. Yuan, M. Duan, Q. Qian *et al.*, 2013 Rice zinc finger protein DST enhances grain production through controlling *Gn1a/OsCKX2* expression. Proc. Natl. Acad. Sci. USA 110: 3167–3172.
Long, M., E. Betrán, K. Thornton, and W. Wang, 2003 The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. 4: 865–875.
Luo, J., H. Liu, T. Zhou, B. Gu, X. Huang *et al.*, 2013 *An-1* encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. Plant Cell 25: 3360–3376.
Manly, K. F., R. H. Cudmore, and J. M. Meer, 2001 Map Manager QTX, cross-platform software for genetic mapping. Mamm. Genome 12: 930–932.
McCouch, S. R., L. Teytelman, Y. Xu, K. B. Lobos, K. Clare *et al.*, 2002 Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res. 9: 199–207.
Miura, K., M. Ikeda, A. Matsubara, X. J. Song, M. Ito *et al.*, 2010 *OsSPL14* promotes panicle branching and higher grain productivity in rice. Nat. Genet. 42: 545–549.
Sayah, D. M., E. Sokolskaja, L. Berthoux, and J. Luban, 2004 Cyclophilin A retrotransposition into *TRIM5* explains owl monkey resistance to HIV-1. Nature 430: 569–573.
Sun, C. Q., X. K. Wang, Z. C. Li, A. Yoshimura, and N. Iwata, 2001 Comparison of the genetic diversity of common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (*O. sativa* L.) using RFLP markers. Theor. Appl. Genet. 102: 157–162.
Tan, L., F. Liu, W. Xue, G. Wang, S. Ye *et al.*, 2007 Development of *Oryza rufipogon* and *O. sativa* introgression lines and assessment for yield-related quantitative trait loci. J. Integr. Plant Biol. 49: 871–884.
Tan, L., X. Li, F. Liu, X. Sun, C. Li *et al.*, 2008 Control of a key transition from prostrate to erect growth in rice domestication. Nat. Genet. 40: 1360–1364.
Tautz, D., and T. Domazet-Lošo, 2011 The evolutionary origin of orphan genes. Nat. Rev. Genet. 12: 692–702.
Turcotte, K., S. Srinivasan, and T. Bureau, 2001 Survey of transposable elements from rice genomic sequences. Plant J. 25: 169–179.
Vaughan, D. A., H. Morishima, and K. Kadowaki, 2003 Diversity in the *Oryza* genus. Curr. Opin. Plant Biol. 6: 139–146.
Wang, Y. H., and J. Y. Li, 2011 Branching in rice. Curr. Opin. Plant Biol. 14: 94–99.
Wang, W., F. G. Brunet, E. Nevo, and M. Long, 2002 Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 99: 4448–4453.
Wang, W., H. Zheng, C. Fan, J. Li, J. Shi *et al.*, 2006 High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791–1802.
Wang, Z., Y. Zou, X. Li, Q. Zhang, L. Chen *et al.*, 2006 Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. Plant Cell 18: 676–687.
Xing, Y. Z., and Q. F. Zhang, 2010 Genetic and molecular bases of rice yield. Annu. Rev. Plant Biol. 61: 421–442.
Xue, W., Y. Xing, X. Weng, Y. Zhao, W. Tang *et al.*, 2008 Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. Nat. Genet. 40: 761–767.
Yamagishi, M., Y. Takeuchi, I. Kono, and M. Yano, 2002 QTL analysis for panicle characteristics in temperate japonica rice. Euphytica 128: 219–224.
Yan, W., P. Wang, H. Chen, H. J. Zhou, Q. P. Li *et al.*, 2011 A major QTL, *Ghd8*, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. Mol. Plant 4: 319–330.
Yoshida, A., M. Sasao, N. Yasuno, K. Takagi, Y. Daimon *et al.*, 2013 *TAWAWA1*, a regulator of rice inflorescence architecture, functions through the suppression of meristem phase transition. Proc. Natl. Acad. Sci. USA 110: 767–772.
Yoshioka, K., W. Moeder, H. G. Kang, P. Kachroo, K. Masmoudi *et al.*, 2006 The chimeric arabidopsis *CYCLIC NUCLEOTIDE-GATED ION CHANNEL11/12* activates multiple pathogen resistance responses. Plant Cell 18: 747–763.
Zhang, Q. F., 2007 Strategies for developing green super rice. Proc. Natl. Acad. Sci. USA 104: 16402–16409.
Zhang, C., J. Wang, N. C. Marowsky, M. Long, R. A. Wing *et al.*, 2013 High occurrence of functional new chimeric genes in survey of rice chromosome 3 short arm genome sequences. Genome Biol. Evol. 5: 1038–1048.
Zhao, L., L. Tan, Z. Zhu, L. Xiao, D. Xie *et al.*, 2015 *PAY1* improves plant architecture and enhances grain yield in rice. Plant J. 83: 528–536.
Zhou, Y., J. Zhu, Z. Li, C. Yi, J. Liu *et al.*, 2009 Deletion in a quantitative trait gene *qPE9–1* associated with panicle erectness improves plant architecture during rice domestication. Genetics 183: 315–324.

Communicating editor: A. Charcosset