


RESEARCH ARTICLE

Open Access



# Genome sequencing of rice subspecies and genetic analysis of recombinant lines reveals regional yield- and quality-associated loci

Xiukun Li<sup>1†</sup>, Lian Wu<sup>1†</sup>, Jiahong Wang<sup>2</sup>, Jian Sun<sup>1</sup>, Xiuhong Xia<sup>1</sup>, Xin Geng<sup>1</sup>, Xuhong Wang<sup>1</sup>, Zhengjin Xu<sup>1</sup> and Quan Xu<sup>1\*</sup> 

## Abstract

**Background:** Two of the most widely cultivated rice strains are *Oryza sativa indica* and *O. sativa japonica*, and understanding the genetic basis of their agronomic traits is of importance for crop production. These two species are highly distinct in terms of geographical distribution and morphological traits. However, the relationship among genetic background, ecological conditions, and agronomic traits is unclear.

**Results:** In this study, we performed the de novo assembly of a high-quality genome of SN265, a cultivar that is extensively cultivated as a backbone *japonica* parent in northern China, using single-molecule sequencing. Recombinant inbred lines (RILs) derived from a cross between SN265 and R99 (*indica*) were re-sequenced and cultivated in three distinct ecological conditions. We identify 79 QTLs related to 15 agronomic traits. We found that several genes underwent functional alterations when the ecological conditions were changed, and some alleles exhibited contracted responses to different genetic backgrounds. We validated the involvement of one candidate gene, *DEP1*, in determining panicle length, using CRISPR/Cas9 gene editing.

**Conclusions:** This study provides information on the suitable environmental conditions, and genetic background, for functional genes in rice breeding. Moreover, the public availability of the reference genome of northern *japonica* SN265 provides a valuable resource for plant biologists and the genetic improvement of crops.

**Keywords:** *Oryza sativa*, De novo assembly, QTL dissection, Yield and quality

## Background

Rice is one of the most important staple crops in the world and provides more than 20% of the calorie intake for more than half of the world's population. Given continuing population growth and increasing competition for arable land between food and energy crops, food security is becoming an ever more serious global problem. Two major types of *Oryza sativa japonica* and *O. sativa indica* subspecies have historically been recognized. Varied degrees of geographical distribution and morphology characters exist between the two subspecies. Elucidation of the relationship among the

functional genomic of *indica* and *japonica*, ecological conditions, and agronomic traits may significantly contribute to the improvement of rice production. China established a nationwide mega project entitled "Breeding and cultivation system of super rice in China" in 1996. After nearly a decade of cultivation, super rice accounts for more than 60% of the total area under rice cultivation and has contributed an estimated two billion dollars to the Chinese national economy [1, 2]. Shennong 265 (SN265), the first released commercial super rice variety, showed not only erect panicles but also strong root activity and high yield in a range of growing environments. SN265 leads the breeding direction as the backbone parent in northern China. Rice genetics and functional genomics have been rapidly advancing, particularly over the last decade, since the first

\* Correspondence: [kobexu34@live.cn](mailto:kobexu34@live.cn)

<sup>†</sup>Xiukun Li and Lian Wu contributed equally to this work.

<sup>1</sup>Rice Research Institute of Shenyang Agricultural University, Shenyang 110866, China

Full list of author information is available at the end of the article



determination of the Nipponbare genome sequence [3]. To improve understanding of the genetic mechanism of hybrid super rice, the genomes of two elite *indica* rice varieties, namely, Zhenshan 97 and Minghui 63, have been assembled [4]. Recently, a near-complete *indica* rice genome of R498 was published, which enriched the implications for plant biology and crop genetic improvement in *indica* [5]. As the *japonica* varieties in northern China have varying degrees of *indica* pedigree introgression [2], the establishment of a reference for *japonica* in northern China is imperative. Thus, the de novo assembly of the SN265 genome will serve as a reference for the discovery of genes and structural variations that contribute to the increase in rice production in super rice varieties in northern China.

Here, we constructed a high-density linkage map by re-sequencing the recombinant inbred lines (RILs) derived from a cross between the *japonica* variety SN265 and *indica* variety R99. We de novo assembled the two parental genomes of SN265 and R99 based on single-molecule real-time sequencing (SMRT) and high-throughput next-generation sequencing (NGS). The RILs were planted in three areas with distinct ecological conditions, and 15 important agronomic traits were investigated. The re-sequencing and assembly of the parental genomes facilitated QTL analysis and candidate gene identification. The influence of genetic background and ecological condition to gene function was investigated in this study.

## Results

### Population sequencing and linkage map construction

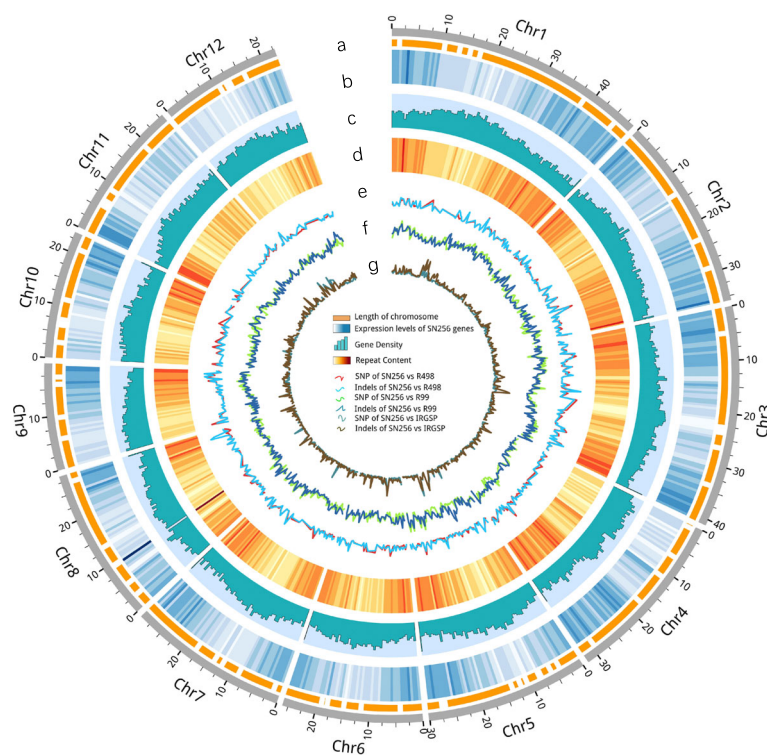
In order to construct the linkage map, the RILs derived from the cross between SN265 and R99, along with the parents, were sequenced on an Illumina HiSeq2500 platform. Through the high-throughput sequencing, we obtained a total of 434.37 Gb of clean data, with approximately 6.25-fold depth for each RILs. For parent lines, 30.0-fold depth and 32.0-fold depth data were generated for R99 and SN265, respectively. We aligned these data to the Os-Nipponbare-Reference-IRGSP-1.0 (<http://rapdb.dna.affrc.go.jp/download/irgsp1.html>) using SOAP2 [6, 7]. Totally 1,708,775 single nucleotide polymorphisms (SNPs) between SN265 and R99 were identified using SOAPSnp [8]. To avoid ambiguity in the analysis, we removed the SNPs that has low genotyping scores or located in highly repetitive regions. As the low-coverage sequencing caused the missing genotype for RILs, the k-nearest neighbor algorithm was used to impute the missing genotypes of each RILs [9]. Subsequently, a recombinant bin map was constructed by 1,456,445 high-quality SNPs. The map contained 3569 recombinant blocks, with the average length of 58.17 kb (Additional file 1: Figure S1 and Additional file 2: Figure S2).

### Assembly of the parental genome

To fill the gaps in the high-quality genome of super rice of northern China, de novo genome assembly of the parent line, SN265, was performed using RIL populations, real-time sequence (SMRT), high-throughput NGS, and RNA-seq. DNA libraries for SMRT sequencing were constructed as described elsewhere [10]. We generated five SMRT cells using P5-C3 SMRT cell chemistry. A total of 24.94 Gb (60-fold) clean data of subread bases with a mean read length of 9.6 kb were generated after filtering the low-quality and short reads. The high-throughput NGS was performed to polish the assembly. DNA libraries for NGS were constructed with size of 270 bp, a total 22.05 Gb clean data (59.66-fold) had quality scores higher than Q20. A 364.45-Mb SN265 genome with a contig N50 value of 6.96 Mb was assembled (Fig. 1). Only 159 erroneous bases, accounting for 0.0000419% of the contig length, were found. A total of 166.68 Mb of repeat sequence was predicted based on a combination of the primary database constructed in this study and Repbase (Additional file 3: Table S1). In addition, a total of 37,609 genes were obtained by de novo prediction, homologous prediction, and RNA-seq analysis after removing repeat sequences. Finally, 95.17% of the genes were functionally annotated with the NR, KOG, GO, KEGG, and TrEMBL databases (Additional file 4: Table S2 and Additional file 5: Table S3). The predicted motifs, non-coding RNAs, and pseudogenes are shown in Additional file 6: Table S4. We further compared the SN265 genome to the current Os-Nipponbare-Reference-IRGSP-1.0 (<http://rapdb.dna.affrc.go.jp/download/irgsp1.html>) [7], and SN265 was 18.3 Mb shorter, but had fewer gaps than the Nipponbare genome. SN265 has longer chromosomes than Nipponbare (i.e., Chr.1, Chr.3, Chr.5, and Chr.8) (Table 1). As the reference genome of typical *indica* rice varieties such as R498, Zhenshan 97, and Minghui 63 was recently released [4, 5], we only conducted a lower-fold SMRT to assemble the R99 genome. A total of 13 Gb (30-fold) of clean data with a mean read length of 8.85 kb were generated. We assembled a 389.6-Mb genome with a contig N50 value of 3.05 Mb. Approximately 663 erroneous bases, accounting for 0.000196% of the contig length, were detected. A total of 166.68 Mb of repeat sequences were predicted, and 36,089 genes were obtained (Additional file 3: Table S1). Finally, 99.92% of the genes were functionally annotated with the NR, KOG, GO, KEGG, and TrEMBL databases (Additional file 4: Tables S2 and Additional file 5: Table S3).

### *Indica* pedigree percentage affects yield and quality traits

To elucidate the relationship between *indica*/*japonica* genetic background and agronomic traits, we determined the *indica* pedigree percentage of each RIL using sequence data. The *indica* pedigree percentage was defined as the ratio of the number of *indica*-type SNPs to



**Fig. 1** Overview of the SN265 reference genome. Tracks from inner to outer circles indicate the following: a contigs and gaps. b expression level genes, c gene density, d repeat sequence content, e SNPs and indels between SN265 and the *indica* (R498) reference genome, f SNPs and indels between SN265 and the *indica* (R99) reference genome, g SNPs and indels between SN265 and the *japonica* (Nipponbare) reference genome

all of the subspecies-specific SNPs for each RILs. The subspecies-specific SNPs were those of the same type in all *japonica*, but not in *indica*, which is based on the divergence of the 517 rice landraces [9]. In total, 100,529 subspecies-specific SNPs were selected. We matched the 1,794,441 SNPs between SN265 and R99 to 100,529

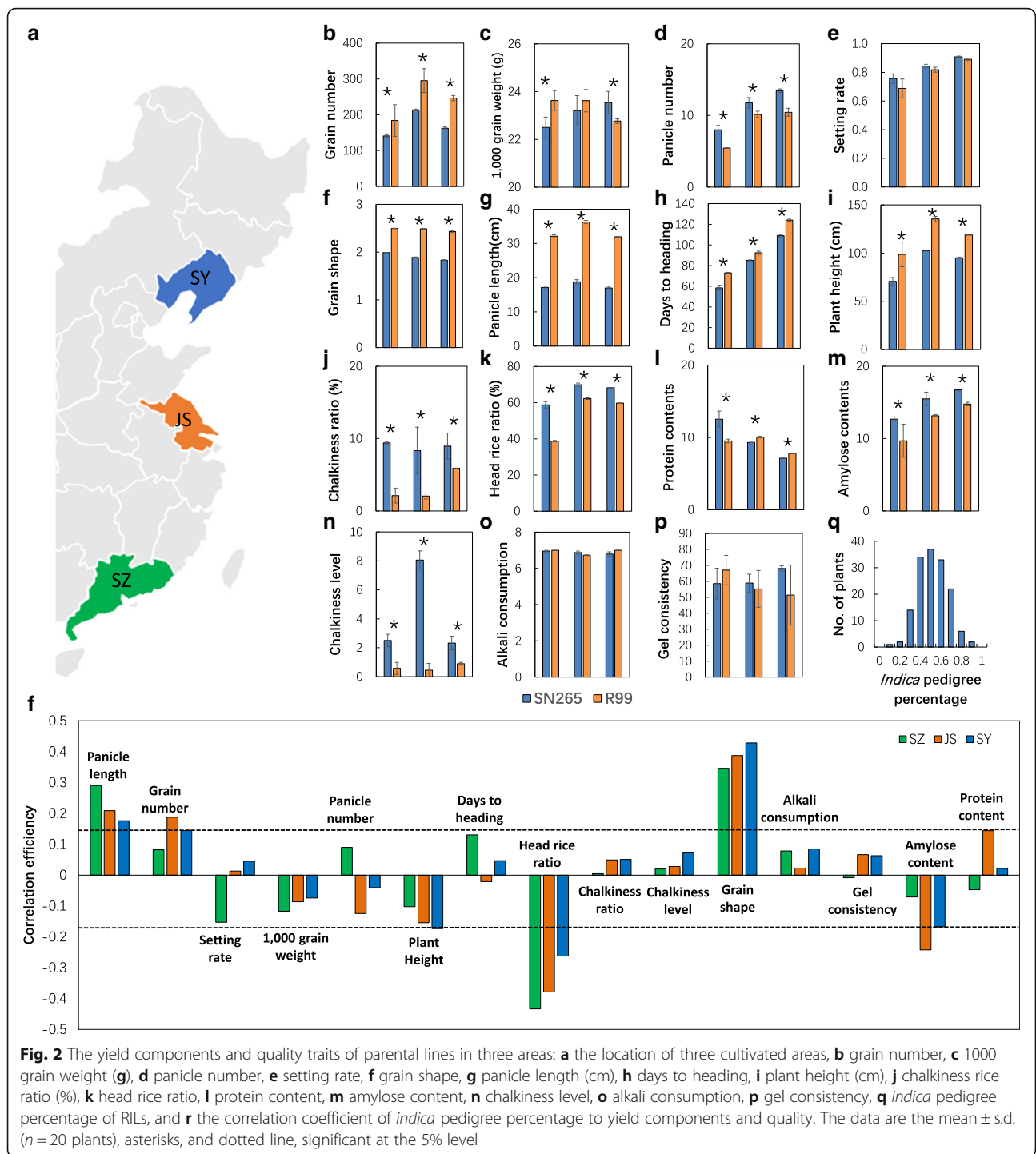
**Table 1** Comparison of basic sequence statistic of SN265 and Os-Nipponbare-Reference-IRGSP-1.0

Chr.	SN265 length	Nip length	SN265 gaps	Nipponbare gaps
Chr1	48,728,733	45,038,628	7	8
Chr2	35,824,122	36,792,250	2	5
Chr3	41,114,166	37,312,367	4	8
Chr4	32,130,185	35,923,694	5	9
Chr5	30,181,130	30,073,434	9	5
Chr6	26,064,998	32,124,787	5	4
Chr7	29,637,245	30,324,621	5	3
Chr8	28,695,623	28,530,022	6	3
Chr9	19,884,354	23,895,720	4	7
Chr10	23,682,709	23,880,549	6	9
Chr11	25,682,812	31,198,810	6	6
Chr12	22,821,698	27,676,856	4	5
Total	364,447,775	382,771,738	63	72

subspecies-specific SNPs, and 61,920 SNPs were merged. The 61,920 SNPs were then used for *indica* pedigree percentage analysis. The results showed that the *indica* pedigree percentage of the RILs followed a normal distribution (Fig. 2). We conducted a correlation analysis of *indica* pedigree percentage to the yield and quality traits. The results showed that the *indica* pedigree percentage showed a significant positive correlation to panicle length and grain shape (the ratio of grain length to grain width) and a negative correlation to head rice ratio in all three of the areas. With increasing latitude, the correlation efficiency between *indica* pedigree percentage and grain shape became larger, and the correlation efficiency of *indica* pedigree percentage to panicle length and head rice ratio became smaller. In JS, the *indica* pedigree percentage also has a significant positive correlation to grain number and a significant negative correlation to amylose content (Fig. 2). As the grain shape always had a significant negative correlation to head rice ratio, we concluded that the *indica* pedigree percentage mainly affects panicle length and grain shape.

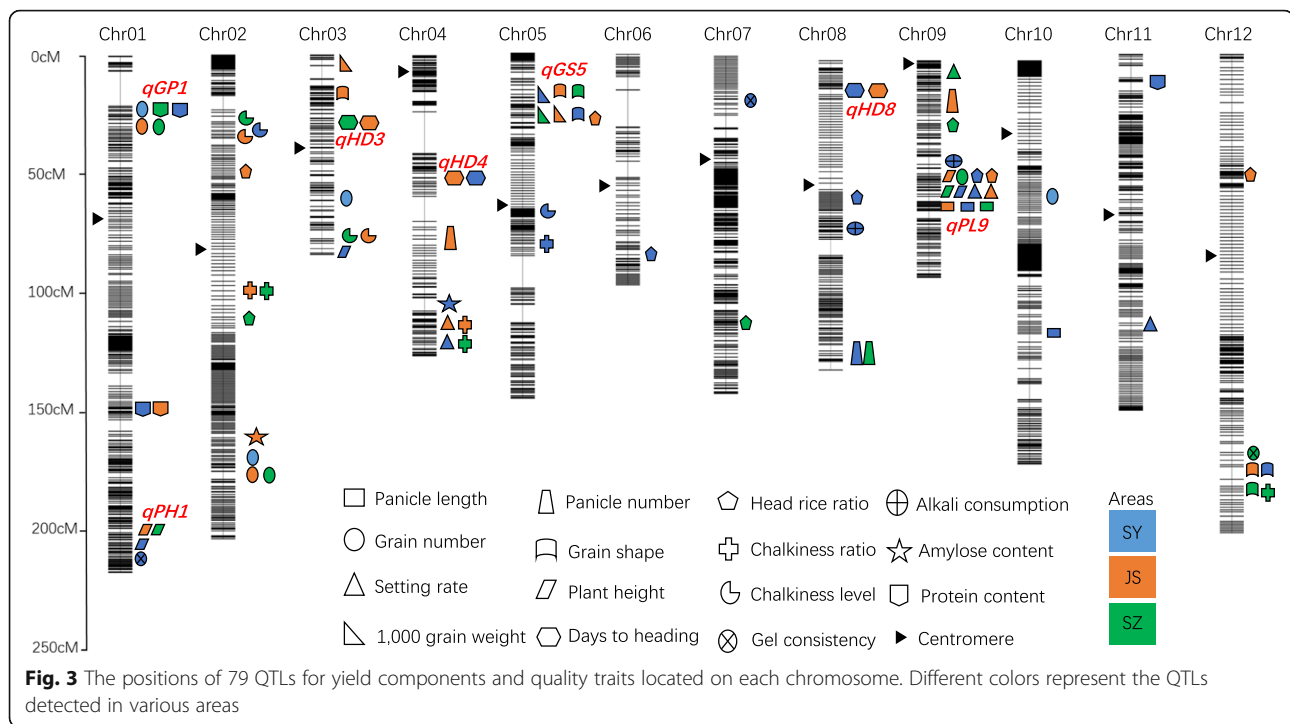
**QTL detection and analysis using the RIL population**

To elucidate the genetic mechanism underlying yield and quality, we primarily focused on 15 traits (panicle length, panicle number, heading rice ratio, alkali consumption,



grain number, grain shape, chalkiness ratio, amylose content, setting rate, plant height, chalkiness level, protein content, 1,000-grain weight, days to heading, and gel consistency) known to be important for rice yield and quality. Both SN265 and R99 showed significant differences in all of the traits except for setting rate, alkali consumption, and gel consistency (Fig. 2). Based on sequence variations in SN265 and R99 and the variant sequences

among the RILs, 1,456,445 SNPs loci were used for linkage analysis. The loci that co-segregated with one another were anchored to the same blocks and designated as “bins.” A total of 3569 bins were used to construct a molecular linkage map using Highmaps software. The phenotype datasets for the RILs collected in the three areas were used for QTL analysis. A total 79 QTLs for all of the traits were mapped independently on rice chromosomes 1 to 12



(Fig. 3 and Additional file 7: Table S5). Among the QTLs, several QTL clusters are highlighted: one locus for grain number per panicle, setting rate, and alkali consumption on the short arm of chromosome 1; one locus for plant height and gel consistency on the long arm of chromosome 1; one locus for grain shape, 1000-grain weight, chalkiness ratio, and alkali consumption on the short arm of chromosome 5; and one locus for plant height, panicle length, grain number per panicle head, and rice ratio on the long arm of chromosome 9. These results suggest that these QTL clusters may be controlled either by one gene with pleiotropy or by a group of closely linked genes. We also found that some loci can be detected in all three of the areas, whereas others can only be detected in only one or two areas.

#### Fine-mapping of QTLs for QTL clustering

To identify candidate genes for the QTL cluster, we conducted fine-mapping using the genome assembly of the two parental lines. We first focused on the *qPL9* cluster on chromosome 9, and the panicle length data in SY was used for fine-mapping. The candidate gene was mapped to a 43-kb interval in block 19948 (Fig. 4). Seven annotated genes were present in this bin. We compared the sequence of seven annotated genes between SN265 and R99 using the assembled genome, revealed a replacement of a 637-bp stretch in the middle region of exon 5 by a 12-bp sequence in SN265 at *DENSE AND ERECT PANICLE 1 (DEP1)* locus (Fig. 4). *DEP1* has been previously shown to be a pleiotropic

major QTL responsible for grain number and panicle architecture [11]. Then, we conducted a co-segregation analysis of *DEP1*, which showed that plants harboring the SN265-type allele of *dep1* develop shorter panicles (Fig. 4). To confirm that *DEP1* is a candidate gene for *qPL9*, we conducted a mutagenesis assay involving *DEP1* in the *japonica* cultivar Sasanishiki using the CRISPR/Cas9 technology. After introducing the construct into rice embryogenic calli by *Agrobacterium*-mediated transformation, we obtained at least 30 independent regenerated transgenic lines. The mutant lines were grown in the field, and the  $T_2$  mutants were sequenced. We detected five homozygous mutants, and the sequencing results are presented in Fig. 4. Among the five mutant lines, four lines harbored a deletion within exon 5, which is predicted to result in a frame shift. One line carried a sequence substitution that was predicted to lead to an amino acid change. We investigated the panicle length in these mutant lines at the mature stage. The results showed that the panicle length in the four lines that harbored a frame shift mutation had significantly decreased compared to Sasanishiki, whereas the substitution line showed a similar panicle length to that in Sasanishiki (Fig. 4). Next, we identified the candidate gene of *qGP1* cluster on the short arm of chromosome 1. The number of grains in SY was used to conduct fine-mapping. The candidate gene was fixed in *Gn1a* [12], and sequence analysis showed that one G/C SNP and a 6-bp deletion occurred in the first exon of R99 compared to the sequence of SN265 (Additional file 8: Figure S3). Then, we

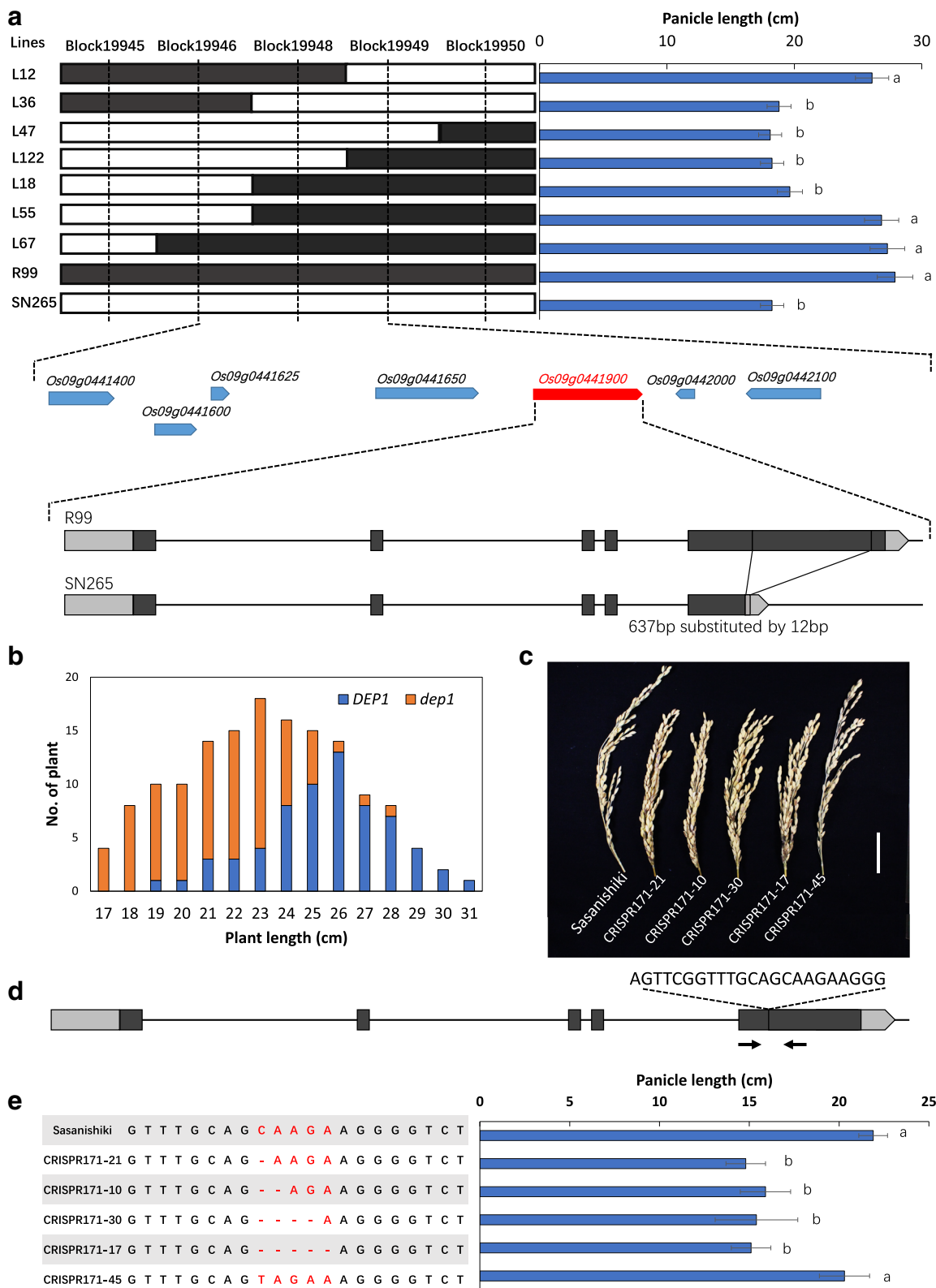


Fig. 4 (See legend on next page.)

(See figure on previous page.)

**Fig. 4** Fine-mapping and CRISPR/Cas9 gene editing of *qPL9/DEP1*. **a** Fine-mapping and sequence comparison of *qPL9*. **b** Co-separation of *DEP1* in RILs, the upper and lower letters indicate the R99-type and SN265-type alleles, respectively. **c** The panicle length of CRISPR/Cas9 gene editing plants and Sasanishiki (WT). **d** Schematic diagram of the genomic region of *qPL9* and the sgRNA target site. **e** Sequence alignment of the sgRNA target region showing altered bases in different mutant lines and the panicle length of mutant lines. The data are the mean  $\pm$  s.d. ( $n = 20$  plants), and the scale bar is 4 cm

identified *SD1* as the candidate gene that corresponds to plant height on the long arm of chromosome 1 [13], and the sequence of first exon significantly differed between R99 and SN265. We identified *GW5* as the candidate gene for grain shape, 1000 grain weight, chalkiness ratio, and alkali consumption and localized this to the short arm of chromosome 5 [14, 15]. A 1,212-bp deletion on the 5.7-kb upstream region of *GW5* was detected in SN265 relative to that in R99 (Additional file 9: Figure S4). We also detected in *qGS12* on chromosome 12, which explains 15.44% of the observed variation. QTL analysis mapped *qGS12* to Block7839, which includes a 23.74-kb region. There were three putative genes within this region, and sequence analysis only detected a single SNP in the third exon of *Os12g0610600* (Additional file 10: Figure S5). *Os12g0610600* was reported as a NAC transcription factor that negatively regulated drought tolerance in rice [16]. The NAC transcription factor family is involved in a number of biological processes in rice, such as drought and salt tolerance, resistance to bacterial leaf blight, heading time, and ABA biosynthesis [17]. The grain shape exhibited significant differences among the combinations between *GW5* and *qGS12* (Fig. 5). Our results show that NAC transcription factors may participate in the regulation of grain shape. However, additional complementary testing as well as supporting genetic evidence is warranted. Moreover, we found that *DTH8*, *SDG708*, and *phytochrome B* (*PHYB*) were the candidate genes for the heading date QTL on chromosomes 8, 4, and 3, respectively [18–20]. There were two SNPs at the first exon region between SN265 and R99 in the *PHYB* locus, a T/G SNP and 19-bp deletion on the *DTH8* locus in R99 and multiple changes leading to a truncated gene structure at the 3' terminal of *SDG708* in R99. The sequence differences in these candidate genes between SN265 and R99 are shown in Additional file 9: Figure S4.

#### The influence of ecological conditions and genetic background to gene function

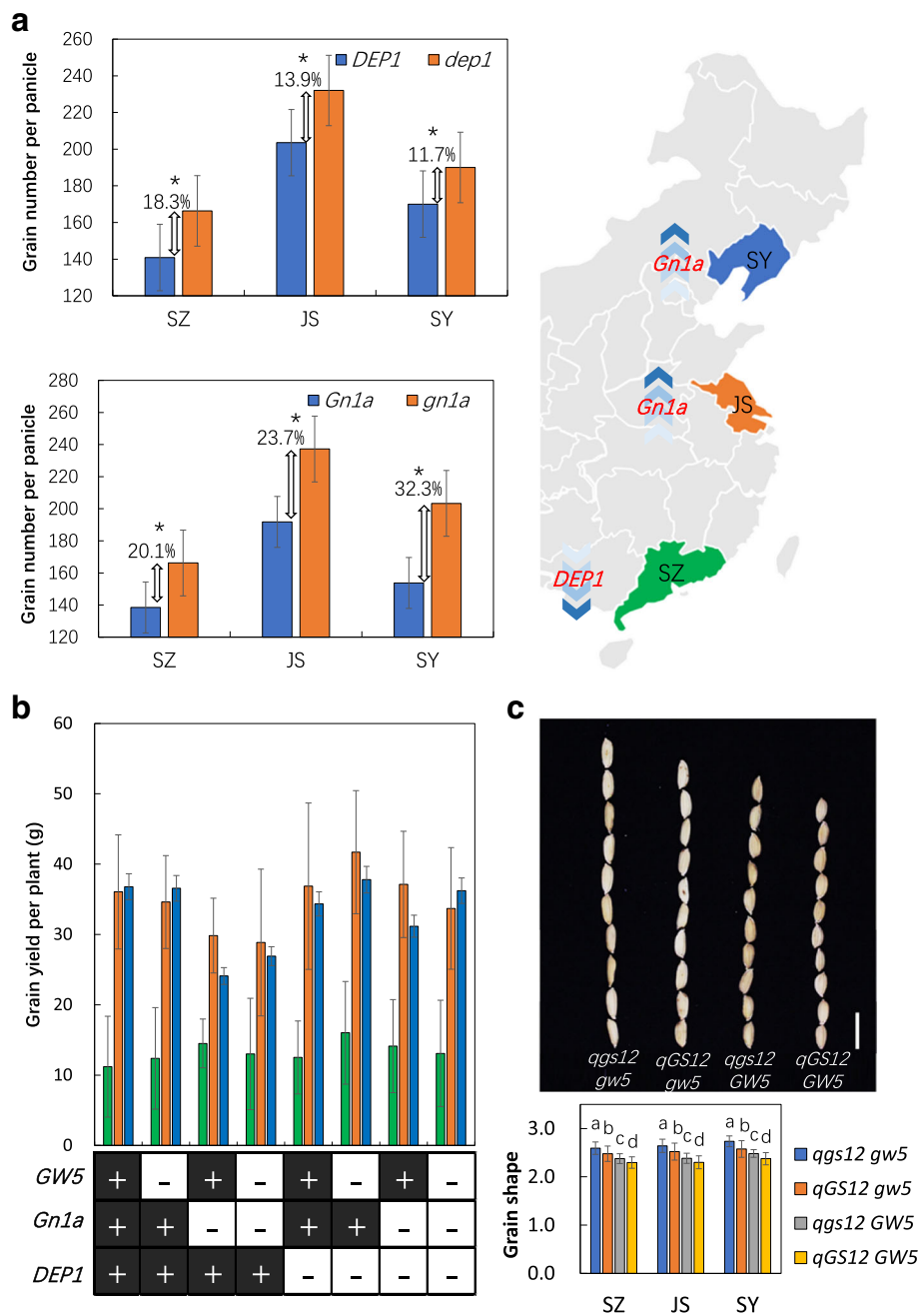
As the RILs were derived from the cross between *indica* and *japonica* and we cultivated the RILs into three areas with distinct ecological conditions, we were able to elucidate the influence of ecological conditions and genetic background to gene function. First, we found that *Gn1a* largely contributed to grain number per panicle in SY and JS, but not in SZ. However, QTL analysis only detected *DEP1* in SZ as a grain number per panicle QTL,

but not in SY and JS. We further compared the plant carrying *DEP1/dep1* and *Gn1a/gn1a* in the three areas, which showed that with increasing latitude, the effect of *dep1* on increasing grain number became weaker, whereas the effect of *gn1a* on increasing grain number became stronger (Fig. 5). Similarly, QTL analysis can detect *DTH8*, *SDG708*, and *PHYB* in JS, but only *PHYB* in SZ, and only *SDG708* and *DTH8* in SY. We selected different gene combinations of *DTH8*, *SDG708*, and *PHYB* in the RILs and compared the heading date of these lines in the three areas. The results confirmed that *DTH8* and *SDG708* barely affect heading date in SZ, and the *PHYB* imparted weaker effects on SY (Additional file 11: Figure S6). In summary, the function of *Gn1a*, *SDG708*, *SD1*, and *GW5* was enlarged with increasing latitude. On the other hand, a disruption in the function of *PHYB* and *DEP1* was observed with increasing latitude. Interestingly, *DTH8* showed the strongest function in middle latitudes and became weaker with increasing or decreasing latitude.

Then, we analyzed the influence of *indica* pedigree percentage to gene function. The RILs were divided into three groups based on the *indica* pedigree percentage: *japonica* group (*indica* pedigree percentage 0~0.4), inter-group (*indica* pedigree percentage 0.4~0.6), and *indica* group (*indica* pedigree percentage 0.6~1). We found that gene function also differs with genetic background. Interestingly, the *japonica*-type allele of *SD1* induces a decrease in plant height with increasing *indica* pedigree percentage, whereas the *indica*-type *sd1* allele increases plant height with higher *indica* pedigree percentage (Additional file 12: Figure S7). Moreover, the *japonica*-type allele of *PHYB* delays the heading date with increasing *indica* pedigree in SZ, but accelerates heading date in JS and SY. Additionally, several gene functions in the inter-type group showed worst agronomic traits among three groups. For example, the inter-group exhibited the lowest grain number per panicle under a *DEP1* and *Gn1a* genetic backgrounds (Additional file 13: Table S6). This may explain why the rare commercial varieties have the inter-type genetic background.

#### Discussion

Rice is a short-day plant model that flowers more rapidly in short-day conditions exhibit delayed flowering under long-day conditions, thereby indicating the existence of critical day length responses [20]. In the wild-type



**Fig. 5** The effects of *DEP1* and *Gn1a* in RILs at low-, middle-, and high-latitude areas. **a** The effect of *DEP1* and *Gn1a* to grain number per panicle in different areas. Asterisks are significant at the 5% level. **b** The effect of the combination among *GW5*, *Gn1a*, and *DEP1* on grain yield per plant in the three areas. **c** The effect of the *qGS12* and *GW5* on grain shape in three areas. Plus sign and lowercase letters indicate the *indica*-type (R99) allele; the minus sign and uppercase letters indicate the *japonica*-type (SN265) allele. The data are the mean  $\pm$  s.d. ( $n = 20$  plants), and the scale bar is 1 cm

plants, although *Hd3a* mRNA is highly expressed at day lengths  $\leq 13$  h, its expression markedly decreased to about one-tenth of the expression, at a day length of 13.5 h and became undetectable at day lengths of  $\geq 14$  h [21]. Our previous study has monitored day length and the temperature during the whole growth season in all

three areas in 2016 [22]. We observed RIL growth under long-day conditions for the whole growing season in SY and JS, and the day length even exceeded 15 h in SY at July. However, the entire growth period of the RILs involved short-day conditions in SZ (Additional file 14: Figure S8). Thus, we assumed that the photoperiod-sensitive



genes play supporting roles in SZ as it maintains optimal day length possibly for the entire growth season, and other factors, such as temperature, play a lead role in establishing heading date. Then, we analyzed the relationship between temperature and heading date in SZ using the method described in our previous study [22]. We found that temperature has a strong correlation with heading date, and this correlation exhibited a positive/negative change by a 10-day rhythm (Additional file 14: Figure S8). Ambient temperature regulates various aspects of plant growth and development, but the actual indicators in rice remain elusive. In *Arabidopsis*, in addition to its photoreceptor function, *PHYB* acts as a temperature sensor [23]. As *PHYB* was the only detected QTL corresponding to heading date in SZ, we assumed that *PHYB* may also be involved in the temperature responses in rice.

Grain weight is a major determinant of crop grain yield and is controlled by naturally occurring quantitative traits loci. Grain shape largely differed between *indica* and *japonica*. *GW5* was detected in most *japonica* cultivars during rice domestication, and a 1212-bp deletion was associated with the increased grain width in *japonica* cultivars [24]. The present study confirmed that the 1212-bp deletion 5.7-kb upstream of *GW5* was the major factor affecting grain width, which explains 38.26% of the observed variation. The SN265-type truncated *dep1* allele is widely distributed among *japonica* varieties in Northeast China and the Yangtze River area [25]. Moreover, the function of the *indica*-type *gn1a* is enhanced with increasing latitude. These results suggest that the introduction of the *japonica* elite allele into the *indica* genetic background or cultivated zone may improve the agronomic traits of *indica* and vice versa. The erect panicle architecture caused by the *dep1* allele significantly increases grain yield; however, the quality traits of these varieties are only considered to be mediocre. As grain width is always significantly positively correlated to chalkiness level, the combination of the *indica*-type *gw5* and *dep1* alleles can simultaneously improve grain yield and morphological traits. Moreover, genotypic analysis of QTLs demonstrated that the haplotype status in RIL lines is responsible for the corresponding traits, whereas the combination of favorable QTLs contributes to relatively high yield per plant. The combination of the SN265-type allele of *DEP1*, *GW5*, and *indica*-type of *Gn1a* is associated with the highest grain yield per plant in all three of the areas (Fig. 5).

With the application of high-throughput sequencing technology, numerous rice accessions have been re-sequenced and phenotyped in the past few years, allowing the exploration of genomic diversity, particularly in terms of identifying loci that are responsive to domestication, as well as in elucidating the molecular mechanism underlying important agronomic traits [9, 26–29]. The de novo assembly of the rice genome provides us with more information to comprehensively capture the genomic diversity in this species [30]. In

this study, we performed de novo assembly of a 364.45-Mb SN265 genome as a reference for super rice in northern China using an RIL population, real-time sequencing (SMRT), and high-throughput NGS.

## Conclusions

Our study identified 79 QTLs that are related to the 15 agronomic traits in three areas with distinct ecological condition and found that several genes underwent functional alterations when the ecological conditions and genetic background were altered. We de novo assembly a super rice variety SN265, and the availability of high-quality reference genomes for the *japonica* subspecies not only facilitates the identification of genes corresponding to agronomic traits but also provides a range of implications for plant biology and crop genetic improvement.

## Methods

### Plant materials and quality measurements

We conducted a cross between “Shennong265” (*Oryza sativa japonica*) and “R99” (*O. sativa indica*) and used the single-seed descendant method to generate RILs with at least 10 generation inbred. A total of 151 RILs were constructed and were used in this study. Field experiments were conducted in three typical rice cultivated areas: the Agricultural Genomics Institute at Shenzhen (SZ; N22°, E114°), the sub-base of China National Hybrid Rice R&D Center in Jiangsu Province (JS; N32°, E120°), and the Rice Research Institute of Shenyang Agricultural University (SY; N41°, E123°) for two growing seasons during 2015–2016. The cultivation method and field management were described in our previous report [22]. We harvested the field examination plants at 45 days after heading for each line in each of the three areas. A total of 20 plants from the middle rows were harvested for each line. The quality measurement was conducted as described in our previous study [22]. We only used the 2016 data in the present study, as the 2 years of data showed similar trends and are highly correlated. All samples were analyzed with two biological replicates.

### DNA extraction and re-sequencing

We sampled the young leaves for each lines 2 weeks after transplanting. To obtain the high-quality DNA, the cetyltrimethylammonium bromide (CTAB) method was used to extract genomic DNA. The sequencing libraries were constructed on the Illumina HiSeq2500 following the manufacturer’s instructions. We aligned the sequencing data to the *japonica* reference genome (Nipponbare, <http://rapdb.dna.affrc.go.jp/download/irgsp1.html/>) using SOAP2 [6]. To construct the genetic map, we combined the co-segregating markers (SNP and/or InDel) into bins using HighMap software [31]. The constructed map

contained 3569 bins, and there were average 247 bins on each chromosome. The map contained 1965.33 cM genetic distance. There were 12 linkage groups in the linkage map, which correspond to the 12 rice chromosomes. The full collinearity between the genetic map and the rice genome was observed, as the minimum value of spearman coefficient for chromosome was 0.9725 (Chr. 6).

#### Single-molecule real-time sequencing (SMRT) and high-throughput NGS

The genomic DNA of each line was extracted from fresh leaves using DNeasy Plant Mini kits (Qiagen, Germany) according to manufacturer's instructions. DNA libraries for SMRT sequencing were performed as described elsewhere [10]. The single-molecule sequencing (SMS) data are assembled following a hierarchical approach: (1) select a subset of longer reads as seed data and correct through canu/falcon [32], (2) use the error-corrected reads for a draft assembly by different assemblers, and (3) polish the draft assembly using Quiver/Arrow and Pilon. In the correction approach, Canu first selects longer seed reads with the settings "genomeSize = 1000000000" and "corOutCoverage = 80," then detects raw reads overlapping through high-sensitive overlapper MHAP (mhap-2.1.2, option "corMhapSensitivity = normal"), then finally performs an error correction through falcon\_sense method (option "correctedErrorRate = 0.025"). In the next approach, with the default parameters, error-corrected reads are trimmed unsupported bases and hairpin adaptors to get the longest supported range. In the last approach, Canu generates the draft assembly by the longest 80 coverage trimmed reads. The draft assembly is polished to obtain the final assembly. Two rounds of polishing are conducted. The first round polishing adopts arrow algorithm by SMS data with the 40 threads, and the second polishing adopts pilon algorithm (v1.22, available at <https://github.com/broadinstitute/pilon>) using illumina data with the parameters "--mindepth 10 --changes --threads 4 --fix bases."

#### Gene annotation

The RNAs of SN265 and R99 were isolated from the fresh leaves using a TaKaRa MiniBEST Universal RNA Extraction Kit according to manufacturer's protocol. The sequencing was performed using the Illumina HiSeq 2500 platform according to manufacturer's instructions. We obtained 8 Gb of RNA-seq data. The MITE-Hunter, LTR\_FINDER v1.05, RepeatScout v1.0.5, and PILER-DF v2.4 were used to construct a primary repeat sequence database using structural prediction and ab initio prediction theory [33–36]. We classified the primary database based on PASTE Classifier and then combined with the Repbase database to form the final repeat sequence

database for the final prediction through Repeat Masker v4.0.6 [37–39]. In protein-coding gene prediction, the repeat elements were masked and excluded from the genome assembly. Gene annotation was performed by three prediction steps: (1) ab initio prediction using Augustus v2.4, Genscan, GlimmerHMM v3.0.4, GeneID v1.4, and SNAP (version 2006-07-28); (2) homologous species prediction based on *Oryza sativa*, *Arabidopsis thaliana*, *Setaria italica*, *Sorghum bicolor*, and *Zea mays* using GeMoMa v1.3.1; and (3) unigene prediction based on full-length transcriptome data assembly with no reference genome was conducted through PASA v2.0.2 [40–45]. The three predictions were integrated through EVM v1.1.1, and final modifications were performed by PASA v2.0.2 [46]. Non-coding RNAs (microRNAs, rRNAs, and tRNAs) were identified with different strategies according to their unique structural features. The miRBase, Rfam, and tRNAscan-SE v1.3.1 databases were used to predict microRNA, rRNA, and tRNA, respectively [47, 48]. We predicted pseudogenes through scanning for homologous genes and excluding genuine genes by GenBlastA v1.0.4 [49]. We selected the candidate genes with premature stop codons and frameshift mutations as the final pseudogene predictions by GeneWise v2.4.1 [50]. In order to annotate genes' function, we blasted the predicted genes to the NR, KOG, GO, TrEMBL, and KEGG databases by BLAST v2.2.31 (-evaluate 1e-5) [50–55]. In addition, the motifs were annotated according to the sequence alignments with the HAMAP, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, SUPERFAMILY, PIRSE, CATH-Gene3D, and PANTHER databases by InterProScan software [56].

#### Vector construction and plant transformation

To conduct the CRISPR/Cas9 gene editing, we performed the vector construction as described by Li et al. [57]. The targeting sequence including PAM sequence (23 bp) was selected in the 5th exon of *DEP1* gene. We confirmed the specificity of targeting sequence by BLAST searching against the rice genome (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) [58]. We performed rice transformation as described elsewhere [59]. We extracted the genomic DNA from transformants, and the genomic DNA were sequenced for mutant identification. The PCR products (200–500 bp) were sequenced and identified using the Degenerate Sequence Decoding method [60].

#### Additional files

**Additional file 1: Figure S1.** Graphic representation of the genotypes of 151 RILs that were identified using a sliding window approach along each chromosome. Different colors represent different genotypes: red, R99; blue, SN265; yellow, heterozygous blocks. (PDF 23 kb)

**Additional file 2: Figure S2.** Collinearity between the bin map derived from the RIL population and the reference genome (Nipponbare). The horizontal and the vertical axes represent the genetic position of the 12 linkage groups from the RIL population map and the physical positions of the 12 rice chromosomes, respectively. The scattered points represent the bin markers used in QTL mapping. The data for each chromosome and linkage group pair are the Spearman correlation coefficient values, which indicated good collinearity when it approaches 1. (PDF 100 kb)

**Additional file 3: Table S1.** Repeat sequence prediction statistics in SN265 and R99. (XLSX 12 kb)

**Additional file 4: Table S2.** Gene prediction statistics of SN265 and R99. (XLSX 9 kb)

**Additional file 5: Table S3.** Gene annotation statistics of SN265 and R99. (XLSX 9 kb)

**Additional file 6: Table S4.** Predicted non-coding RNAs, pseudogenes, and motifs. (XLSX 9 kb)

**Additional file 7: Table S5.** The yield components and quality traits of *japonica*, intermediate type and *indica*-type plants in the three areas. (XLSX 14 kb)

**Additional file 8: Figure S3.** Fine-mapping and sequence comparison of *qGP1(Gn1a)*: a, the *qGP1* was mapped between to Block244; b, the annotated genes in Block244; and c, the sequence difference of *Gn1a* between SN265 and R99. (PDF 103 kb)

**Additional file 9: Figure S4.** Sequence comparison of *PHYB*, *SD1*, *DTH8*, *SDG708*, and *GW5* between SN265 and R99. (PDF 47 kb)

**Additional file 10: Figure S5.** Candidate gene prediction of grain shape regulate locus on chromosome 12. (PDF 23 kb)

**Additional file 11: Figure S6.** The effects of *DTH8*, *SDG708*, and *PHYB* in different areas: a, the heading date of the different combination of *DTH8*, *SDG708*, and *PHYB*; b, the major effect heading gene in three areas. The data are the mean  $\pm$  s.d. ( $n = 20$  plants), "+" and "-" indicate the R99-type and S265-type alleles, respectively. (PDF 185 kb)

**Additional file 12: Figure S7.** The influence of ecological conditions and genetic background on gene function. The uppercase and lowercase lettered gene name indicates R99-type and SN265-type alleles, respectively. (PDF 378 kb)

**Additional file 13: Table S6.** The QTL analysis for 15 agronomic traits in the three areas. (XLSX 12 kb)

**Additional file 14: Figure S8.** The heading time of RILs in the three areas: a, the heading data of RILs and the day length in three areas; b, the correlation coefficient between air temperature and heading time. (PDF 109 kb)

#### Funding

The National Natural Science Foundation of China (31430062 and 31501284) supported this study.

#### Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. The sequences reported in this paper have been deposited in the National Center for Biotechnology Information (NCBI). This Whole Genome Shotgun data has been deposited at DDBJ/ENA/GenBank under the accessions QWGC00000000 (SN265) and QWGD00000000 (R99). PRJNA486237 and PRJNA486425 (the *Oryza sativa* raw sequence reads), and SRP158741 contain the raw sequence reads of RILs. The seeds of recombinant inbred line populations and the parents are available from the corresponding author on reasonable request.

#### Authors' contributions

ZX and QX designed this study and contributed to the original concept of the project. XL and LW performed most of the experiments. JW and JS participated the genome assemble. XG and XX participated in the assessment of yield components in three areas. XW participated in quality measurement. QX wrote the paper. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### 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.

#### Author details

<sup>1</sup>Rice Research Institute of Shenyang Agricultural University, Shenyang 110866, China. <sup>2</sup>Biomarker Technologies Corporation, Beijing 101300, China.

Received: 4 July 2018 Accepted: 6 September 2018

Published online: 18 September 2018

#### References

- Qian Q, Guo L, Smith SM, Li J. Breeding high-yield superior quality hybrid super rice by rational design. *Natl Sci Rev*. 2016;3(3):283–94.
- Sun J, Liu D, Wang J-Y, Ma D-R, Tang L, Gao H, Xu Z-J, Chen W-F. The contribution of intersubspecific hybridization to the breeding of super-high-yielding japonica rice in northeast China. *Theor Appl Genet*. 2012;125(6):1149–57.
- Fujisawa M, Baba T, Nagamura Y, Nagasaki H, Waki K, Vuong H, Matsumoto T, Wu JZ, Kanamori H, Katayose Y. The map-based sequence of the rice genome. *Nature*. 2005;436(7052):793–800.
- Zhang J, Chen LL, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song JM, Xie W. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc Natl Acad Sci U S A*. 2016;113(35):E5163.
- Du H, Ying Y, Ma Y, Qiang G, Cao Y, Zhuo C, Ma B, Ming Q, Yan L, Zhao X. Sequencing and de novo assembly of a near complete indica rice genome. *Nat Commun*. 2017;8:15324.
- Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25(15):1966–7.
- Kawahara Y, Bastide MDL, Hamilton JP, Kanamori H, Mccombie WR, Shu O, Schwartz DC, Tanaka T, Wu J, Zhou S. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 2013;6(1):4.
- Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. SNP detection for massively parallel whole-genome resequencing. *Genome Res*. 2009;19(6):1124.
- Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet*. 2010;42(11):961–7.
- Hammer M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods*. 2015;12(8):780–6.
- Huang X, Qian Q, Liu Z, Sun H, He S, Luo D, Xia G, Chu C, Li J, Fu X. Natural variation at the DEP1 locus enhances grain yield in rice. *Nat Genet*. 2009;41(4):494–7.
- Ashikari M, Sakakibara H, Lin S, Yamamoto T, Takashi T, Nishimura A, Angeles ER, Qian Q, Kitano H, Matsuoka M. Cytokinin oxidase regulates rice grain production. *Science*. 2005;309(5735):741–5.
- Sasaki A, Ashikari M, Ueguchi-Tanaka M, Itoh H, Nishimura A, Swapan D, Ishiyama K, Saito T, Kobayashi M, Khush GS. A mutant gibberellin-synthesis gene in rice. *Nature*. 2002;416(6882):701–2.
- Weng J, Gu S, Wan X, Gao H, Guo T, Su N, Lei C, Zhang X, Cheng Z, Guo X. Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res*. 2008;18(12):1199–209.
- Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M. Deletion in a gene associated with grain size increased yields during rice domestication. *Nat Genet*. 2008;40(8):1023–8.
- Fang Y, Xie K, Xiong L. Conserved miR164-targeted NAC genes negatively regulate drought resistance in rice. *J Exp Bot*. 2014;65(8):2119–35.

17. Nuruzzaman M, Manimekalai R, Sharoni AM, Satoh K, Kondoh H, Ooka H, Kikuchi S. Genome-wide analysis of NAC transcription factor family in rice. *Gene*. 2010;465(1):30–44.
18. Wei X, Xu J, Guo H, Jiang L, Chen S, Yu C, Zhou Z, Hu P, Zhai H, Wan J. DTH8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. *Plant Physiol*. 2010;153(4):1747–58.
19. Liu B, Wei G, Shi J, Jin J, Shen T, Ni T, Shen WH, Yu Y, Dong A. SET DOMAIN GROUP 708, a histone H3 lysine 36-specific methyltransferase, controls flowering time in rice (*Oryza sativa*). *New Phytol*. 2016;210(2):577–88.
20. Ishikawa R, Aoki M, Kurotani K, Yokoi S, Shinomura T, Takano M, Shimamoto K. Phytochrome B regulates heading date 1 (Hd1)-mediated expression of rice florigen Hd3a and critical day length in rice. *Mol Gen Genomics*. 2011; 285(6):461–70.
21. Itoh H, Nonoue Y, Yano M, Izawa T. A pair of floral regulators sets critical day length for Hd3a florigen expression in rice. *Nat Genet*. 2010;42(7):635–8.
22. Li X, Wu L, Geng X, Xia X, Wang X, Xu Z, Xu Q. Deciphering the environmental impacts on rice quality for different rice cultivated areas. *Rice*. 2018;11(1):7.
23. Jung JH, Domijan M, Klose C, Biswas S, Ezer D, Gao M, Khattak AK, Box MS, Charoensawan V, Cortijo S. Phytochromes function as thermosensors in *Arabidopsis*. *Science*. 2016;354(6314):886–9.
24. Ashikari M, Wu J, Yano M, Sasaki T, Yoshimura A. Rice gibberellin-insensitive dwarf mutant gene Dwarf 1 encodes the alpha-subunit of GTP-binding protein. *Proc Natl Acad Sci U S A*. 1999;96(18):10284–9.
25. Xu H, Zhao M, Zhang Q, Xu Z, Xu Q. The DENSE AND ERECT PANICLE 1 (DEP1) gene offering the potential in the breeding of high-yielding rice. *Breed Sci*. 2016;66(5):659–67.
26. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*. 2018;557(7703):43.
27. Huang X, Yang S, Gong J, Zhao Q, Feng Q, Zhan Q, Zhao Y, Li W, Cheng B, Xia J. Genomic architecture of heterosis for yield traits in rice. *Nature*. 2016; 537(7622):629–33.
28. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W. A map of rice genome variation reveals the origin of cultivated rice. *Nature*. 2012;490(7421):497–501.
29. Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet*. 2012;44(1):32.
30. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet*. 2018;50(2):278.
31. Liu D, Ma C, Hong W, Huang L, Liu M, Liu H, Zeng H, Deng D, Xin H, Song J. Construction and analysis of high-density linkage map using high-throughput sequencing data. *PLoS One*. 2014;9(6):e98855.
32. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27(5):722.
33. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics*. 2005;21(suppl\_1):i351.
34. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*. 2010;38(22):e199.
35. Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 2007;35(Web Server issue):W265–8.
36. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. *Bioinformatics*. 2005;21(Suppl 1):i152.
37. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capi P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2009;10(4):276.
38. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 2005;110(1–4):462–7.
39. Tarailogrovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2004; Chapter 4(Unit 4):Unit 4.10.
40. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003;19(suppl\_2):215–25.
41. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 2004;20(16):2878–9.
42. Blanco E, Parra G, Guigó R. Using geneid to identify genes. *Current Protocols in Bioinformatics*. 2007;18(1):Unit 4.3.
43. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5(1):59.
44. Jens K, Michael W, Erickson JL, Schattat MH, Jan G, Frank H. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res*. 2016;44(9):e89.
45. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics*. 2006;7(1):327.
46. Haas BJ, Salzberg SL, Wei Z, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 2008;9(1):R7.
47. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. 2005;33:121–4.
48. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29(22):2933–5.
49. She R, Chu JS, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res*. 2009;19(1):143–9.
50. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res*. 2004;14(5):988–95.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
52. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;27(1):29–34.
53. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*. 2001;29(1):22–8.
54. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I. The Swiss-Prot knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. 2003;31(1):365–70.
55. Marchlerbauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, Deweescott C, Fong JH, Geer LY, Geer RC, Gonzales NR. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res*. 2011;39:225–9.
56. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001;17(9):847–8.
57. Li W, Zhu Z, Chern M, Yin J, Yang C, Ran L, Cheng M, He M, Wang K, Wang J. A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell*. 2017;170(1):114–26.
58. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol*. 2013;31(9):827–32.
59. Nishimura A, Aichi I, Matsuoka M. A protocol for agrobacterium-mediated transformation in rice. *Nat Protoc*. 2006;1(6):2796.
60. Ma X, Zhang Q, Zhu Q, Liu W, Chen Y, Qiu R, Wang B, Yang Z, Li H, Lin Y. A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol Plant*. 2015;8(8):1274–84.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

