Bulked Segregant Analysis Coupled With Whole-Genome Sequencing (BSA-Seq) and Identification of a Novel Locus, qGL3.5, That Regulates Grain Length

Haiyan Zhao
South China Agricultural University

Yuebin Zheng
South China Agricultural University

Feng Bai
South China Agricultural University

Yang Liu
South China Agricultural University

Sicheng Deng
South China Agricultural University

Xiangdong Liu
South China Agricultural University

Lan Wang ( wanglan@scau.edu.cn )
South China Agricultural University https://orcid.org/0000-0003-2242-1028

Research Article

Keywords: Rice, Grain length, BSA, Gene cloning

DOI: https://doi.org/10.21203/rs.3.rs-263682/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Rice grain length (GL) directly affects the yield and quality of this species. Very few GL-related genes cloned are applied in production because their yield-increasing effect was not obvious. In this study, the two bulk-DNA pools (L-pool and S-pool) and their parents' (KJ01 and Huaye 4) DNAs were subjected to high-throughput sequencing. After assessing the quality of the data, we obtained a total of 100.22 Gb of high-quality data; the average coverage depth was 55x, and the genome coverage was 96.51%. After combining the association results of the ED and SNP index methods, we mapped the GL genes to a 0.34 Mb “hotspot” region on chromosome 3, which contains 37 genes related to various traits. The 37 predicted genes were further analyzed by the use of the Gene GO, COG database and so on. Thirty-three genes were annotated by GO functions. According to the GO annotations, three genes whose molecular function involved in the plasma membrane and intracellular membrane-bound organelles were detected via CRISPR/ Cas9 editing technology. ORF33 was verified to regulate GL and was the target gene qGL3.5. These results provides a new gene resource for rice grain shape breeding and a starting point for functional characterization of the wild rice GL gene.

Introduction

Rice is among the most important staple crop species and feeds approximately one-third of the global population. Rice grain yield is determined by three main characteristics: grain weight, number of grains per panicle, and number of panicles per plant (Li et al. 2018). In terms of grain weight, the critical factor that affects yield is grain size, which includes grain length (GL), grain width, grain thickness, and the degree of filling (Huang et al. 2013). Among these four factors, GL contributes the most to grain weight (Lin and Wu, 2003), and previous studies have shown that thin grains have better quality (Chen et al. 1997; Shi and Zhu 1997). In addition, GL is also an important factor of rice quality.

GL is a complex quantitative trait that is controlled by 2~3 or more genes (Mckenize and Rutger 1983). To date, at least 102 QTLs controlling GL have been located (Huang et al. 2013). However, only eleven genes have been cloned and conducted further functional studies. The genetic mechanism underlying the effects of these genes has been elucidated and is very complex, and their modes of regulation are also diverse. qGL3/GL3.1 encodes a PPKL family serine/threonine phosphatase and restricts cell division in spikelet hulls by dephosphorylating cyclin T1;3 (Cyct1;3) (Qi et al. 2012; Zhang et al. 2012). GS3, a major gene controlling GL and grain weight, encodes a putative Gγ protein and inhibits interactions between the Gβ protein RGB1 (G protein γ subunit 1) and two other Gγ proteins to regulate grain size (Sun et al. 2018). GLW7 encodes the plant-specific transcription factor OsSPL13, which positively regulates GL and yield (Si et al. 2016). GL7 regulates GL by affecting the expression of two linked genes (Wang et al. 2015). GL6 encodes a plant AT-rich sequence and zinc-binding (PLATZ) transcription factor; this protein has a premature stop codon and reduces the expression or causes the loss of function of GL6, affecting rice GL (Wang et al. 2019).
Although some genes regulating GL have been cloned, few genes have been applied in production, and the overall regulatory networks underlying the associated processes remain poorly understood. With the development of high-throughput sequencing in recent years, high-efficiency, low-cost whole-genome sequencing has become available. Approaches based on bulked segregant analysis (BSA) coupled with whole-genome sequencing (BSA-seq) have been widely applied for mapping several important agronomic-related genes in rice (Zhang et al. 2019), and gene cloning and functional research have become easier. BSA was used to construct two pools by mixing the DNA of individuals with extreme traits in a segregating population (Takagi et al. 2013). By analyzing the differences between single-nucleotide polymorphism (SNPs) and insertions/deletions (InDels) between the two pools, we can quickly locate the molecular markers tightly linked to a target gene (Klein et al. 2018). This method is appropriate for qualitative traits controlled by one gene and quantitative traits controlled by major QTLs. Currently, BSA-seq, a rapid approach for identifying QTLs, is widely used in rice (Yang et al. 2013; Tao et al. 2018; Gao et al. 2019; Liang et al. 2020), maize (Klein et al. 2018), and cotton (Rodriguez-Uribe et al. 2011).

Interestingly, we obtained a wild rice inbred line Huaye 4 (Oryza rufipogon Griff) that presented very short GL—only 7.03 mm. An F$_2$ population was then generated from a cross of the cultivar KJ01 with a long GL of 13.33 cm and Huaye 4. QTL genetic linkage analysis showed that there was a novel gene that controlled GL and explained 54.85% of the phenotypic variation (Zheng et al. 2020). In this study, BSA-seq was used to narrow the gap of 300 kb on chromosome 3 by the use of an F$_2$ population comprising 1,307 individuals, and no reported GL (GL) genes were found in this interval, which included 37 candidate genes according to annotation analysis of SNP variation. Among these genes, ORF33, namely, qGL3.5, was found via clustered, regularly interspaced, short palindromic repeat (CRISPR)/CRISPR-associated 9 (Cas9) to regulate GL. Our study provides a new gene resource for rice grain shape breeding and a starting point for functional characterization of the wild rice GL gene.

**Materials And Methods**

**Plant materials**

The wild rice inbred line Huaye 4, cultivar KJ01, and their F$_2$ progeny were used. These materials were grown in the field at South China Agricultural University during the late season of 2015. Ten rows were planted in each of the test plots, and the row spacing was 18 cm×21 cm.

**Measurements of characteristics**

The rice grains were harvested at the mature stage by collecting three major panicles of each plant, after which the materials were put into marking envelopes. Samples of 30 randomly selected grains were assessed for GL, and the length of 10 grains placed end to end was measured using a ruler with a precision of ± 0.01 mm. The grains of each plant were replicated three times. The average length of a single rice grain was then calculated.

**Construction of bulked DNA pools for sequencing**
Two bulk DNA pools for sequencing were first constructed by selecting individuals exhibiting an extreme phenotype from among 1,307 F$_2$ segregating plants. The individuals that produced the top 50 longest grains were grouped into the L-pool, and the individuals that produced the top 50 shortest grains were grouped into the S-pool from among the F$_2$ segregating lines.

Total genomic DNA was extracted from young healthy leaves of both parents and 100 selected F$_2$ segregating individuals via the sodium dodecyl sulfate (SDS) procedure and purified by RNase A. The DNA concentration and quality were estimated with a Nanodrop 2000 UV-Vis spectrophotometer (NanoDrop, Wilmington, DE, USA), and the DNA concentration was diluted to 3 $\mu$g/$\mu$L; the total DNA was not less than 150 $\mu$g. The individual genomic DNA of the 50 L-pool individuals was mixed together equally, constituting the long-seed bulk DNA pool (L-pool), and the individual genomic DNA of the 50 S-pool individuals was mixed together equally, constituting the short-seed bulk DNA pool (S-pool). The genomic DNA of the two bulk DNA pools and both parents was prepared for subsequent high-throughput sequencing.

**Sequencing data analysis and comparison**

The four samples were subjected to high-throughput sequencing on an Illumina HiSeq™ 2500 instrument (Illumina, Inc., San Diego, CA, USA) at Biomarker Technologies Corporation in Beijing. Real-time monitoring was performed for each cycle during sequencing, and the percentage of high-quality reads with quality scores greater than Q30 (a quality score of 30, indicating a 0.1% chance of an error and thus 99.9% confidence) among the raw reads and the GC content were calculated for quality control. Afterward, the sequencing data were compared with the sequence of the reference genome of Nipponbare (download link: http://rapdblegacy.dna.affrc.go.jp/).

**SNP screening**

The detection of SNPs was mainly performed via the GATK software toolkit (McKenna et al. 2010). According to the positioning results of clean reads in the reference genome, Picard (http://sourceforge.net/projects/picard/) was used for preprocessing, such as marking duplicates, and GATK was used for local realignment and base recalibration to ensure the accuracy of detected SNPs. GATK was then used for SNP detection and filtering, after which a final set of SNPs were obtained.

**Association analysis**

The relative marker abundance in the L-pool was calculated as the number of reads of the maternal allele divided by the number of reads of the paternal allele, whereas in the S-pool, the relative marker abundance was calculated as the number of reads of the paternal allele divided by that of the maternal allele. It was expected that the larger the relative abundance was, the greater the possibility that the marker was associated with GL. SNP index association analysis (Abe et al. 2012) and ED association analysis (Deza et al. 2009) were used in this research, and in the present study, L refers to the KJ01 parent, S refers to the parent of Huaye 4, aa represents the L-pool and ab represents the S-pool.
**SNP index association analysis**

The idea behind SNP index association analysis was recently published; this analysis is a type of method used to calculate genotype frequency differences between two bulks satisfied by $\Delta$(SNP index). A relatively close marker is associated with traits, while a relatively close $\Delta$(SNP index) is associated with the value 1. The $\Delta$(SNP index) was calculated as follows: Laa is the depth of the aa group derived from L, while Saa indicates the depth of the aa group derived from S. Lab means the depth of the ab group derived from L, while Sab represents the depth of the ab group derived from S.

\[
\text{SNP index (aa)} = \frac{\text{Laa}}{\text{Laa} + \text{Saa}}
\]

\[
\text{SNP index (ab)} = \frac{\text{Lab}}{\text{Lab} + \text{Sab}}
\]

\[
\Delta \text{SNP index} = \text{SNP index (aa)} - \text{SNP index (ab)}
\]

**ED association analysis**

ED association analysis is a method for calculating ED (the quadratic sum root of differences between bulks based on the depth of the four types of bases) and is satisfied by ED. In theory, the higher the ED value is, the closer the object site.

The ED was calculated as described below. Aaa, Caa, Taa, and Gaa represent the depth of bases A, C, T and G, respectively, at a particular site in the large seed bulk. Aab, Cab, Tab, and Gab represent the depths of bases A, C, T and G, respectively, at a particular site in the small seed bulk.

In this study, high-throughput sequencing combined with BSA was used to detect SNP tags between the two bulked DNA pools and to rapidly identify marker-intensive hotspots associated with GL in the rice genome.

**GO analysis of selected candidate genes**

The GO database is a structured standard biological annotation system that produces standard vocabulary for the functions of genes and their products. The database structure is divided into multiple levels. The lower the level is, the more specific the functions the nodes represent. The candidate genes that were selected were subjected to GO-based functional enrichment analysis, and they were classified according to three categories: biological processes, cellular components and molecular functions.

**CRISPR/Cas9 transgene analysis**

Two targets were designed for the potential candidate genes *ORF14, ORF33* and *ORF25* via the website CRISPR-GE (http://skl.scu.edu.cn/). The expression cassette vector was constructed by multiple rounds of PCR (Ma et al. 2015). The target sequences were introduced into sgRNA expression cassettes by overlapping PCR. The purified PCR products (sgRNA expression cassettes) were inserted into a pYLCRISPR/Cas9 vector based on Golden Gate system. The ligated products harboring the sgRNA
expression cassettes were directly used to transform *E. coli* DH5α competent cells. The CRISPR/Cas9 constructs that were successfully generated were introduced into *Agrobacterium tumefaciens* strain EHA105. *Agrobacterium tumefaciens*-mediated gene transfer experiments were subsequently carried out in the background of Huaye 4. The results were analyzed via the sequential decoding method (http://skl.scau.edu.cn/) to identify positive transgenic plants. Three wild-type plants and twelve mutant plants were selected at maturity to measure their GL.

**Results**

**Phenotypic analysis of both parents and genetic analysis of GL within the F₂ population**

All of the traits of the wild rice inbred line Huaye 4, including plant height and GL, remained stable by continuous self-crossing and selective breeding for several years (Figure 1A). The plant height was approximately 41.20 cm, and the GL was approximately 7.03 mm. Huaye 4 was used as one of the parents and crossed with the rice cultivar KJ01, which has a long GL-13.33 mm. There were significant differences in GL, grain width, plant height, and panicle length between the two parents (Table 1, Figure 1B). Genetic analysis of GL for 1,307 separate F₂ individuals showed that the frequency distribution of GL was continuously distributed, that the means deviated from those of the long-grain parent, and that the variation range of the GL was very large, from 7.07 mm to 13.33 mm, in the F₂ population (Figure 1C). Therefore, the GL of the two parents was a typical quantitative trait controlled by multiple genes.

**Table 1 Phenotypic value of traits for KJ01 and Huaye 4**

|              | Grain length (mm) | Grain width (mm) | Plant height (cm) | Panicle length (cm) |
|--------------|-------------------|------------------|-------------------|--------------------|
| KJ01         | 13.33±0.05        | 3.30±0.23        | 123.26±1.37       | 34.75±0.48         |
| Huaye 4      | 7.03**±0.09       | 1.80**±0.15      | 63.67**±0.44      | 16.63**±0.24       |

"**" indicates a highly significant difference at the P=0.01 level.

**Analysis of the quality of the bulked DNA sequence data**

Fifty individuals that produced extremely long grains (11.57-13.33 mm) and 50 individuals that produced extremely short grains (7.07-8.13 mm) from the F₂ population were selected for preparation of the long-seed pool (L-pool) and short-seed pool (S-pool). In total, 100.60 Gb of raw data were obtained from four samples (KJ01, Huaye 4, the L-pool and the S-pool), and 100.22 Gbp of clean data were attained after filtering. The Q30 percentage was 93.86%, and the guanine-cytosine (GC) content was 45.83%. Of the high-quality data, **55,999,600** reads were obtained from KJ01, and **45,059,434** reads were obtained from Huaye 4. The read numbers for the S-pool and L-pool were **127,409,412** and **106,871,512**, respectively. The average percentage between the sample and the reference genome was 96.64%, the average coverage depth was 55x, and the genome coverage was 96.51% (at least one base coverage). Four
samples (one each of KJ01, Huaye 4, the L-pool and the S-pool), with 16.72 Gb, 13.46 Gb, 31.93 Gb and 38.09 Gb, respectively, were sequenced, and their GC contents were 45.72%, 45.54%, 46.04% and 46.03%, respectively.

Version RGAP7 of the *Oryza sativa japonica* genome, which comprised 374 Mb and had a GC content of 43.55%, was used as the reference genome. The average depth and Cov-ratio-10x (%) of KJ01 and Huaye 4 were 38x (88.40%) and 31x (90.19%), respectively. Similarly, the average depth and Cov-ratio-10x (%) of the L-pool and the S-pool were 70x (95.23%) and 83x (95.89%), respectively (Table S1).

The coverage of the genome can reflect the amount of variation from the reference genome. The more regions are covered, the more mutation sites can be detected. The coverage was influenced mainly by sequencing depth and distance between the sample and reference genome. The depth of genome coverage affects the accuracy of mutation detection, and the accuracy of mutation detection is relatively high in regions with relatively high coverage depths (nonrepetitive sequence regions).

Sequencing quality statistics showed that our sequencing data were of high quality, and good consistency, and could be used for subsequent bioinformatics analysis.

**Detection of SNPs between samples and reference genomes**

SNP mutations can be divided into transition and transversion types. Mutations involving the same type of bases, such as changes between a purine and another purine, and between a pyrimidine and another pyrimidine, are called transitions. Mutations involving different types of bases, such as changes between a purine and pyrimidine, are called transversions. Generally, transitions are more likely to occur than transversions, so the transition/transversion (Ti/Tv) ratio is generally greater than one. For diploid or polyploid species, if a certain SNP site on homologous chromosomes is the same base, the SNP site is referred to as homozygous. If a SNP site on a homologous chromosome involves different types of bases, the SNP site is referred to as heterozygous SNP.

In the present study, the higher the number of homozygous SNPs was, the greater the difference between the sample and the reference genome were, and the greater the number of heterozygous SNPs was, the higher the degree of heterozygosity of the sample was. These results are related to the material used. Compared with the number of SNPs in the reference genome, the number of SNPs in KJ01 and Huaye 4 were 1,443,316 and 1,245,956, respectively. For the two bulked pools, the numbers of SNPs in the S-pool and L-pool were 2,216,645 and 2,136,482, respectively (Table S2). The number of SNPs common to both the L-pool and S-pool was 2,111,496 (Figure 2). Finally, 2,337,510 SNPs were selected for further analysis after two rounds of sequencing and the exclusion of low-quality fragments.

For SNP index association analysis, 5,201 SNPs with multiple mutations were filtered, 772,175 SNPs were the same in the two bulk pools, 23,119 SNPs had a read support of less than four, and 311,069 SNPs were nonexistent in the parents. A total of 1,225,946 SNPs were ultimately selected for further analysis. Similarly, 1,773,999 SNPs were selected for euclidean distance (ED) association analysis.
SNP index and ED association analysis

Two BSA-seq analysis methods, the SNP index and ED, were used to identify the candidate genomic regions associated with GL. After filtering, a total of 1,225,946 SNPs were used for association analysis through the SNP index method. A SNP index of the L-pool and S-pool was calculated for each identified SNP in the genome, and an average SNP index was computed for each 1 Mb interval via a 10 kb sliding window (Figure 3A, 3B). By combining the SNP index information of the L-pool and S-pool, we calculated a \( \Delta (\text{SNP index}) \), and \( \Delta (\text{SNP index}) \) trends were visualized by means of a sliding window (Figure 3C). Using the association threshold of 0.99, we identified only one genomic region distributed on chromosome 3 that was significantly correlated with GL. The length of the region was 25.53 Mb and contained 4,734 genes.

A total of 1,773,999 SNPs were used for association analysis through the ED method. The association threshold was 2.5958 of the ED data, and 4 associated regions on chromosome 3 were closely correlated with the GL trait (Figure 3D). By analyzing the 4 associated regions, we found that the first region was 0.03 Mb in the interval from 23,160,000-23,190,000 bp and contained 6 candidate genes, the second region had only one gene at the 23,210,000 bp position, the third region was 0.11 Mb in the interval from 23,250,000-23,360,000 bp and contained 16 genes, and the fourth region was 0.20 Mb in the interval from 23,830,000-24,030,000 bp and contained 34 genes. The total length of these regions was 0.34 Mb, and they contained a total of 57 genes (Table 2).

By combining the association results of the ED and SNP index methods, we found that the intersections of the genomic regions were consistent with the associated regions identified via the ED method. The 0.34 Mb region of chromosome 3, which contained 57 genes, identified via BSA-seq was the “hotspot” for GL of rice. There were no genes controlling GL reported in these genomic regions.

These results in turn allowed us to narrow the genomic region of the QTL linkage analysis, which identified the major QTLs as being within the interval of markers PSM379-RID24455, spanning 7.4 Mb (Zheng et al, 2020).

Table 2 Association region for GL according to a combined analysis of the ED and SNP index methods.

| Chromosome ID | Start       | End         | Size (Mb) | Gene number |
|---------------|-------------|-------------|-----------|-------------|
| Chr3          | 23,160,000  | 23,190,000  | 0.03      | 6           |
| Chr3          | 23,210,000  | 23,210,000  | 0.00      | 1           |
| Chr3          | 23,250,000  | 23,360,000  | 0.11      | 16          |
| Chr3          | 23,830,000  | 24,030,000  | 0.20      | 34          |
Gene Ontology (GO) enrichment analysis of candidate genes

The different SNPs in the exon regions of the 57 genes between the parents were analyzed. There were a total of 246 nonsynonymous mutation SNPs involving 37 genes. To identify the candidate genes for GL, the 37 predicted open reading frames (ORFs) were queried against the GO, nonredundant protein (NR), SwissProt, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Cluster of Orthologous Groups (COG) of proteins databases.

Of these candidate genes, 37 were annotated in the NR database; 37 were listed in the NT database; 37 were listed in the TrEMBL database; 23 were listed in the SwissProt database; 33 were listed in the GO database; 3 were identified via KEGG pathway analysis; and 9 were functionally annotated in the COG database (Table S3). Afterward, 33 ORFs annotated via GO enrichment were further analyzed, of which 29 were related to cell components, 23 were related to molecular functions, and 18 were related to biological processes (Table 3).

Directed acyclic graphs generated by topGO intuitively show GO terms and their hierarchical relationship with gene enrichment in an associated region. An acyclic graph is a graphical representation of the results of GO enrichment analysis of genes in associated regions; branches represent inclusion relationships, and the functional scope defined from top to bottom becomes increasingly specific (Figure S1). By performing a topGO analysis of the associated regions in cell components, molecular functions, and biological processes, we found eight GO terms associated with the 37 genes, of which five of the terms were related to cell components, two were related to biological processes, and only one was related to molecular functions (Table 4). For these results, the smaller the Kolmogorov-Smirnov (KS) value is, the more significant the enrichment. The plasma membrane had the most significant topGO enrichment (GO:0005886). The three nonsynonymous mutated genes in the associated region, ORF13, ORF14 and ORF19, were involved in plasma membrane enrichment according to the GO annotation.

In addition, ORF33 and ORF25 were involved in intracellular membrane-bound organelles according to the GO annotation (GO:0043231; GO:0044444). The nonsynonymous mutated ORF33 gene was annotated as being associated with extensin in the SwissProt database. Moreover, this gene was found to be related to extensor proteins in tobacco. Studies have shown that extensor proteins are important structural proteins in the plant cell wall and play a role in cell wall strengthening. Studies have also shown that the expression of extensor proteins is negatively correlated with the degree of cell expansion, and increasing the expression of extensor proteins may promote an increase in cell density in local areas of the tissues or organs in which they are expressed (Roberts et al. 2006). ORF25 encodes a protein that participates in carbohydrate transport and metabolism, and is considered a formin-like protein according to the SwissProt database. Among the molecular functions, ATP binding was most significant (GO:005524), and ORF13 and ORF14, two nonsynonymous mutated genes, were also enriched in ATP binding.

Table 3 GO annotations of nonsynonymous mutated genes in the associated region
| Ontology | GO term                        | Annotated gene number |
|----------|--------------------------------|-----------------------|
| C C      | cell                           | 29                    |
| C C      | membrane                       | 3                     |
| C C      | macromolecular complex         | 2                     |
| C C      | organelle                      | 27                    |
| C C      | organelle part                 | 2                     |
| C C      | membrane part                  | 2                     |
| C C      | cell part                      | 29                    |
| C C      | catalytic activity             | 17                    |
| M F      | transporter activity           | 3                     |
| M F      | binding                        | 17                    |
| M F      | enzyme regulator activity      | 1                     |
| B P      | metabolic process              | 17                    |
| B P      | cellular process               | 13                    |
| B P      | single-organism process        | 3                     |
| B P      | response to stimulus           | 3                     |
| B P      | localization                   | 3                     |
| B P      | multiorganism process          | 2                     |
| B P      | biological regulation          | 3                     |
| **Total**| **—**                          | **33**                |

Notes: CC, cellular component; MF, molecular function; BP, biological process.

**Table 4** Top-ranked GO gene annotations in associated regions among 37 genes
| GO ID      | Ontology | Term                          | Significant | Expected | KS      |
|------------|----------|-------------------------------|-------------|----------|---------|
| GO:0005886 | CC       | plasma membrane               | 3           | 2.79     | 1.40E-13|
| GO:0005634 | CC       | nucleus                       | 4           | 4.01     | 2.00E-11|
| GO:0016021 | CC       | integral component of membrane| 2           | 2.4      | 1.70E-09|
| GO:0005773 | CC       | vacuole                       | 1           | 0.99     | 1.40E-05|
| GO:0005789 | CC       | endoplasmic reticulum membrane| 2           | 0.16     | 0.00012 |
| GO:0005524 | MF       | ATP binding                   | 2           | 2.89     | 2.30E-06|
| GO:0006281 | BP       | DNA repair                    | 2           | 0.3      | 0.00134 |
| GO:0006882 | BP       | cellular zinc ion homeostasis | 1           | 0.01     | 5.30E-05|

**Functional verification of candidate genes via CRISPR/Cas9 gene-editing technology**

On the basis of the results of the GO annotation, *ORF14, ORF33* and *ORF25* were selected to verify their molecular function via CRISPR/Cas9 gene-editing technology. Interestingly, after the *ORF33* gene was knocked out in Huaye 4 by CRISPR/Cas9 gene-editing technology, the target gene was inserted such that there was an extra base in the exon, which resulted in a frameshift mutation, and the GL increased by 0.11 compared with that of its wild type in the T0 generation. The edited homozygous T0 seeds were planted, and T1 seeds were retrieved, which were subsequently planted; the genotype and phenotype of T1 were investigated. The mutant phenotype of the T1 plants was stable in the segregating population (Figure 4). The GL of T1 plants became longer if the *ORF33* gene was edited, with or without the presence of the hygromycin gene. These results indicated that *ORF33* regulated GL, and was the *qGL3.5* target gene.

*ORF33* of Huaye 4 was knocked out via CRISPR/Cas9. Shown are the differences in phenotype and genotype between wild-type and mutant plants. A, Plant type comparison between the wild type and mutant; B, GL comparison between the wild type and mutant. C, BLAST-reported sequence of the two target sites.

**Discussion**

**BSA-seq enhances mapping accuracy and saves time**

The high-throughput Illumina sequencing platform can be used to obtain a large amount of data in a short time. It can be used to construct a double terminal sequencing library, and its base accuracy rate
exceeds 98.5% (Metzker 2010). Similarly, BSA is a good time-saving method for the identification of QTLs (Wu et al. 2019).

In this study, a total of 100.22 Gb (clean data) was attained after filtering of four different materials. The average percent similarity between the sample and the reference genome (version RGAP7 of *Oryza sativa japonica*) was 96.64%, the average coverage depth was 55x, and the genome coverage was 96.51%. For the two bulked pools, there were 2,216,645 and 2,136,482 SNPs in the S-pool and L-pool, respectively (Table S2). Ultimately, 1,225,946 SNPs were selected for SNP index analysis, and 1,773,999 SNPs were selected for ED association analysis. By combining the association results of the ED and SNP index methods, we mapped the candidate genes of GL to an interval of 0.34 Mb on chromosome 3 and narrowed the mapping region of a previous QTL genetic linkage analysis (Zheng et al. 2020).

**Wild rice contains many valuable genes**

As research has advanced, the genetic resources derived from cultivated rice have become increasingly narrower. Because it has excellent genes associated with resistance and yield, wild rice has become a good candidate to explore valuable genes to enhance rice productivity (Rahman et al. 2007; Guo et al. 2009). To date, an increasing number of finely mapped or cloned genes from wild rice have been successively reported.

*GAD1*, derived from *Oryza rufipogon*, encodes a small secretory signal peptide and regulates both GL and awn development (Jin et al. 2016). *qHD7.2*, which delays rice flowering and shortens panicle length under long-day conditions, was identified in *Oryza rufipogon* Griff. (Li et al. 2018). Wang et al. (2019) cloned *GL6*, which controls rice GL and is derived from *Oryza rufipogon* W1943; *GL6* encodes a PLATZ transcription factor. *qGCA1* derived from *Oryza longistaminata* significantly improves the general combining ability of plant height, spikelet number, and yield per plant and was delimited to an interval of approximately 72 kb on chromosome 1 (Fan et al. 2019). *ORF12* from *Oryza minuta* was identified as a candidate gene (*TGW12*) that controls rice thousand-grain weight (Li et al. 2020). In our study, by using the BSA method and CRISPR/Cas9 gene-editing technology, we identified a new gene, *qGL3.5*, that controls rice GL and is derived from *Oryza rufipogon*.

**Extensin proteins regulates GL**

*qGL3.5* was related to extensor proteins identified in tobacco research. Extensin is a hydroxyproline-rich protein attached to cellulose filaments in plant cell walls, that functions were in pollen recognition and fertilization (Wu et al. 2001), cell division and differentiation, cell elongation termination (Ito et al. 1998), senescence and shedding (Merkouropoulos et al. 2003), and biotic and abiotic stress responses (Showalter 1993). Ke et al. (2019) reported that *OsPEX1*, a leucine-rich repeat extension-like gene, mediates lignin biosynthesis and/or deposition in rice and leads to reduced plant height and enhanced lodging resistance.
The biological function of most cloned GL genes involves extension, and there is a correlation between cell elongation and the expression of extensor proteins. Previous studies have shown that the expression level of extensors is negatively correlated with the degree of cell extension, and increasing the expression level of extensors may promote an increase in cell density in the local area of tissues or organs in which these genes are expressed (Roberts et al. 2006). Studies on GL genes show that GL7 positively regulates longitudinal cell elongation via tandem duplication of a 17.1 kb segment, increasing the GL (Wang et al. 2015). GS6, on the other hand, encodes a GRAS domain and affects lemma development by regulating the number of cells in glumes, thus negatively regulating the size of rice grains (Sun et al. 2013).

Interactions of genes controlling GL

To date, four genes controlling GL on chromosome 3 have been cloned: GS3, GL3.1, qTGW3 (qGL3.3), and qGL3.5. It is worth considering whether these genes undergo interactions. Previous studies have shown that there are interaction effects on grain shape; some were positive regulatory effects, and others were negative regulatory effects.

GL3.3 epistatically interacts together with GS3 to produce extralong rice grains (Xia et al. 2018). Zhang et al. (2016) studied grain trait variation of 110 lines to evaluate the interactions of qGL3.1 and GS3 and found that mutations in qGL3 and GS3 (via qGL3.1 and GS3 loss-of-function mutations, respectively) significantly affected GL and grain weight, and the genetic effects of qGL3.1 on grain phenotypes were stronger than those of GS3; that is, qGL3.1 and GS3 have an additive effect on the regulation of rice GL. By comparatively analyzing the primary panicle transcriptomes of 93-11, NIL-GS3, NIL-qGL3 and NIL-GS3/qGL3, Gao et al. (2015) found that GS3 and qGL3.1 have additive effects on rice GL regulation. However, other genes controlling grain shape may also undergo interactions. Yan et al. (2011) showed that GS3, GW2 and qSW5 interact, qSW5 and GW2 positively regulate the expression of GS3, and GS3 explains the GL effect of qSW5. Perhaps qGL3.5 interacts with these genes, which needs further verification. It is very important to clearly elucidate the regulatory network of these GL-related genes, which is valuable to rice grain shape breeding.

GL is a complex trait that is usually controlled by multiple genes. Zheng et al. (2020) identified three major QTLs in a gene mapping region. There may be other genes in addition to qGL3.5 that regulate GL in our mapping region. Therefore, the molecular functions of other candidate genes need to be verified.

Conclusion

In this study, BSA-seq was used to rapidly and successfully explore candidate genes controlling rice GL. Fifty individuals that produce extremely long grains and 50 individuals that produce extremely short grains among an F2 population of 1,307 individuals were selected to construct an L-pool and an S-pool. In total, 100.60 Gb of raw data were obtained from four samples (KJ01, Huaye 4, the L-pool and the S-pool), and 100.22 Gbp (clean data) was used after filtering. The average percent similarly between the
sample and the reference genome was 96.64%, the average coverage depth was 55x, and the genome coverage was 96.51%. Four samples (KJ01, Huaye 4, the L-pool and the S-pool) were sequenced, with 16.72 Gb, 13.46 Gb, 31.93 Gb and 38.09 Gb, respectively. Compared with the number of SNPs in the reference genome, 1,225,946 SNPs were selected for SNP index analysis, and 1,773,999 SNPs were selected for ED association analysis. By combining the association results of the ED and SNP index methods, we found that four associated regions on chromosome 3 were closely correlated with the GL trait. In these genomic regions, no genes controlling GL have been reported. The total length of these regions was 0.34 Mb, containing 37 genes. GO enrichment analysis revealed that 33 ORFs were annotated. Among them, three genes whose molecular function involved the plasma membrane and intracellular membrane-bound organelles were detected via CRISPR/Cas9 gene-editing technology. ORF33 was verified to regulate GL and was the qGL3.5 target gene.

**Abbreviations**

GL: grain length; BSA: bulked segregant analysis; BSA-seq: based on bulked segregant analysis coupled with whole-genome sequencing; SNPs: single-nucleotide polymorphisms; ED: euclidean distance; GO: gene ontology; NR: nonredundant protein; KEGG: Kyoto Encyclopedia of Genes and Genomes; COG: Cluster of Orthologous Groups; CRISPR: clustered, regularly interspaced, short palindromic repeat; Cas9: CRISPR-associated 9; PLATZ: plant AT-rich sequence and zinc-binding; L-pool: long-seed pool; S-pool: short-seed pool; ORFs: open reading frames

**Declarations**

**Acknowledgements** This work was supported by Guangdong Provincial Natural Science Foundation Program (No. 2019A1515011826), and Guangdong Provincial Special Projects in Key Field of Rural Revitalization (No. 2020ZDZX1038), and University Students’ Innovation Project of South China Agricultural University (No. S202010564128). We thank Dr Zhixiong Chen, Dr Jinwen Wu, and Dr. Muhammad Qasim Shahid for their assistance.

**Author contributions statement** HYZ, YBZ, FB, YL and SCD performed the experiments. XDL and LW conceived and designed the experiments. HYZ and LW analyzed the data and wrote the paper. XDL and LW supervised and complemented the manuscript. All authors read and approved the final manuscript.

**Compliance with ethical standards**

Conflict of interest The authors declare no conflicts of interest.

**References**

1. Li N, Xu R, Duan P, Li Y. Control of grain size in rice. Plant Reprod, 2018,31:237-251.
2. Huang R Y, Jiang L R, Zheng J S, Wang T S, Wang H C, Huang Y M, Hong Z L. Genetic bases of rice grain shape: so many genes, so little know. Trends in plant science, 2013, 18(4):218-226.
3. Lin L H, Wu W R. Mapping of QTLs underlying grain shape and grain weight in rice. Molecular Plant Breeding, 2003, 1(3): 337-342.

4. Chen N, Luo Y K, Zhu Z W, Zhang B P, Zheng Y C, Xie L H. Correlation between eating quality and physic-chemical properties of high grain quality rice. Chinese J. Rice Sci, 1997, 11(2): 70-76.

5. Shi C H, Zhu J. Analysis of genetic covariances between plant agronomic traits and milling quality traits of indica rice. Journal of Zhejiang Agricultural University, 1997, 23(3): 331-337.

6. Mckenzie K S, Rutger J N. Genetic analysis of amylose content, alkali spreading score, and grain dimensions in rice. Crop Science, 1983, 23(2): 306-313.

7. Qi P, Lin Y S, Song X J, Shen J B, Huang W, Shan J X, Zhu M Z, Jiang L, Gao J P, Lin H X. The novel quantitative trait locus GL3.1 controls rice grain size and yield by regulating Cyclin-T_1;3. Cell Res, 2012, 22: 1666-1680.

8. Zhang X J, Wang J F, Huang J, Lan H X, Wang C L, Yin C F, Wu Y Y, Tang H J, Qian Q, Li J Y, Zhang H S. Rare allele of OsPPKL1 associated with grain length causes extra-large grain and a significant yield increase in rice. Proc Natl Acad Sci USA, 2012, 109: 1240.

9. Sun S S, Wang L, Mao H L, Shao L, Li X H, Xiao J H, Ouyang Y D, Zhang Q F. A G-protein pathway determines grain size in rice. Nat Commun, 2018, 9: 851.

10. Si L Z, Chen J Y, Huang X H, Gong H, Luo J H, Hou Q Q, Zhou T Y, Lu T T, Zhu J J, Shangguan Y Y, Chen E W, Gong C X, Zhao Q, Jing Y F, Liu K Y, Wei X H, An K, An G, Han Ben. OsSPL13 controls grain size in cultivated rice. Nat Genet, 2016, 48: 447-546.

11. Wang Y X, Xiong G S, Hu J, Jiang L, Yu H, Xu J, Fang Y X, Zeng L J, Xu E, Xu J, Ye W J, Meng X B, Liu R F, Chen H Q, Jing Y H, Wang Y H, Zhu X D, Li J Y, Qian Q. Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nature Genetics, 2015, 47: 944-948.

12. Wang A H, Hou Q Q, Si L Z, Huang X H, Luo J H, Lu D F, Zhu J J, Shangguan Y Y, Miao J S, Xie Y F, Wang Y C, Zhao Q, Feng Q, Zhou C C, Li Y, Fan D L, Lu Y Q, Tian Q L, Wang Z X, Han B. The PLATZ Transcription factor GL6 affects grain length and number in rice. Plant Physiology, 2019, 180: 2077-2090.

13. Zhang H W, Wang X, Pan Q C, Li P, Liu Y J, Lu X D, Zhong W S, Li M Q, Han L Q, Li J, Wang P X, Li D D, Liu Y, Li Q, Yang F, Zhang Y M, Wang G Y, Li L. QTG-Seq accelerates QTL fine mapping through QTL partitioning and whole-genome sequencing of bulked segregant samples. Mol. Plant, 2019, 12: 426-437.

14. Takagi H, Abe A, Yoshida K, Natsume S, Mitsuoka C, Uemura A, Utashti H, Tamiru M, Takuno S. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant Journal, 2013, 74(1): 174-183.

15. Klein H, Xiao Y G, Conklin P A, Govindarajulu R, Kelly J A, Scanlon M J, Whipple C J, Bartlett M. Bulked-segregant analysis coupled to whole genome sequencing (BSA-Seq) for rapid gene cloning in Maize. G3: Genes/Genomes/Genetics, 2018, 8: 3583-3592.

16. Yang Z M, Huang D Q, Tang W Q, Zheng Y, Liang K J, Cutler A J, Wu W R. Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled
extremes. PloS One, 2013,8(7):e68433.

17. Tao Y H, Niu Y N, Wang Y, Chen T X, Naveed S A, Zhang J, Xu J L, Li Z K. Genome-wide association mapping of aluminum toxicity tolerance and fine mapping of a candidate gene for Nrat1 in rice. Plos One, 2018,13(6):e0198589.

18. Gao J, Dai G X, Zhou W Y, Liang H F, Huang J, Qing D J, Chen W W, Wu H, Yang X H, Gao L J, Deng G F. Mapping and identifying a candidate gene p lr4, a recessive gene regulating purple leaf in rice, by using bulked segregant and transcriptome analysis with next-generation sequencing. International Journal of Molecular Sciences, 2019, 20:4335.

19. Liang T M, Chi W C, Huang L K, Qu M Y, Zhang S B, Chen Z Q; Chen Z J, Tian D G, Gui Y J, Chen X F, Wang Z H, Tang W Q, Chen S B. Bulked segregant analysis coupled with whole-genome sequencing (BSA-Seq) mapping identifies a novel pi27 haplotype conferring basal resistance to rice blast disease. International Journal of Molecular Sciences, 2020,21:2162.

20. Rodriguez-Uribe L, Higbie S M, Stewart J M, Wilkins T, Lindemann W, Sengupta-Gopalan C, Zhang J F. Identification of salt responsive genes using comparative microarray analysis in Upland cotton (Gossypium hirsutum). Plant Science An International Journal of Experimental Plant Biology. 2011,180:461-469.

21. Zheng Y B, Li Z, Zhao H Y, Zhu G F, Liao Z Y, Zhu Z H, Wang L. Mapping quantitative trait loci associated with grain length and genetic analysis of major quantitative loci in rice. Acta Bot Boreal, 2020,40(4):0598-0604

22. Metzker M L. Sequencing technologies-the next generation. Nature Reviews Genetics, 2010, 11 (1):31-46.

23. Wu S L, Qiu J, Gao Q K. QTL-BSA: a bulked segregant analysis and visualization pipeline. Computational Life Sciences, 2019, 11:730-737.

24. Rahman M L, Chu S H, Choi M S, Qiao Y L, Jiang W Z, Piao R H, Khanam S, Cho Y I, Jeung J U, Jena K, Koh H J. Identification of QTLs for some agronomic traits in rice using an introgression line from Oryza minuta. Mol Cells, 2007,24(1):16-26

25. Guo S B, Qin F L, Zhang D P, Lin X H. Characterization of interspecific hybrids and backcross progenies from a cross between Oryza minuta and Oryza sativa. Science in China Series C-Life Science. 2009, 52:1148-1155

26. Jin J, HUA L, ZHU Z F, et al. GAD1 encodes a secreted peptide that regulates grain number, grain length, and awn development in rice domestication[J]. The Plant Cell, 2016, 28(10): 2453-2463.

27. Li J, Xu R, Wang C C, Qi L, Zheng X M, Wang W S, Ding Y B, Zhang L Z, Wang Y Y, Cheng Y L, Zhang L F, Qiao W H, Yang Q W. A heading date QTL, 2, from wild rice (Oryza rufipogon) delays flowering and shortens panicle length under long-day conditions. Science Reports, 2018,8:2928

28. Fan F F, Long W X, Liu M M, Yuan H R, Pan G J, Li N W, Li S Q. Quantitative Trait Locus Mapping of the Combining Ability for Yield-Related Traits in Wild Rice Oryza longistaminata. J. Agric. Food Chem. 2019, 67, 8766−8772
29. Li X Q, Wei Y, Li J, Yang F W, Chen Y, Chen Y H, Guo S B, Sha A H. Identification of QTL \textit{TGW12} responsible for grain weight in rice based on recombinant inbred line population crossed by wild rice (\textit{Oryza minuta}) introgression line K1561 and indica rice G1025. BMC Genetics, 2020, 21:10.

30. Wu H, Graaf B de, Mariani C, Cheung A Y. Hydroxyproline-rich glycoproteins in plant reproductive tissues: structure, functions and regulation. Cellular and Molecular Life Sciences, 2001, 58(10):1418-1429.

31. Ito M, Kodama H, Komamine A, Watanabe A. Expression of extensin genes is dependent on the stage of the cell cycle and cell proliferation in suspension-cultured catharanthus roseus cells. Plant Molecular Biology, 1998, 36(3):343-351.

32. Merkouropoulos G, Shirsat A H. The unusual \textit{Arabidopsis} extensin gene \textit{atext1} is expressed throughout plant development and is induced by a variety of biotic and abiotic stresses. Planta, 2003, 217(3):356-366.

33. Showalter A M. Structure and function of plant cell wall proteins. Plant Cell, 1993, 5(1):9-23.

34. Ke S W, Luan X, Liang J Y, Hung Y H, Hsieh T F, Zhang X Q. Rice OsPEX1, an extensin-like protein, affects lignin biosynthesis and plant growth. Plant Molecular Biology, 2019, 100:151–161.

35. Roberts K, Shirsat A H. Increased extension levels in \textit{Arabidopsis} affect inflorescence stem thickening and height. Journal of Experimental Botany, 2006, 57(3):537-545.

36. Sun L J, Li X J, Fu Y C, Zhu Z F, Tan L B, Liu F X, Sun X Y, Sun X W, Sun C Q. \textit{GS6}, a member of the \textit{GRAS} gene family, negatively regulates grain size in rice. J Integr Plant Biol, 2013, 55(10):939-949.

37. Xia D, Zhou H, Liu R J, Han W, Li P B, Wu B, Chen J x, Wang L Q, Gao G J, Zhang Q L, He Y Q. 3, a Novel QTL Encoding a GSK3/SHAGGY-like Kinase, Epistatically Interacts with \textit{GS3} to Produce Extra-long Grains in Rice. Molecular plant, 2018, 11(5):754-756.

38. Zhang Y D, Zhu Z, Zhao Q Y, Chen T, Yao S, Zhou L H, Zhao L, Zhao C F, Wang C L. Haplotype types of \textit{qGL3} and their roles in grain size regulation with \textit{GS3} alleles in rice. Genet Mol Res. 2016 15(1): doi: 10.4238/gmr.15017587.

39. Gao X Y, Zhang X Y, Lan H X, Huang J, Wang J F, Zhang H S. The additive effects of \textit{GS3} and \textit{qGL3} on rice grain length regulation revealed by genetic and transcriptome comparisons. BMC Plant Biol. 2015, 15:156.

40. Yan S, Zou G H, Li S J, Wang H, Liu H Q, Zhai G W, Guo P, Song H M, Yan C J, Tao Y Z. Seed size is determined by the combinations of the genes controlling different seed characteristics in rice. Theoretical and Applied Genetics, 2011, 123(7): 1173-1181.

41. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulski K, Kemytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo M A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010, 20:1297-303.

42. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Yoshida K, Mitsuoka C, Tamiru M, Innan H, Cano L, Kamoun S, Terauchi R. Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol. 2012, 30(2):174-178.
43. Deza M M, Deza E. Encyclopedia of Distances, 2\textsuperscript{nd} Springer, Berlin, 2009; pp 94

44. Ma X L, Zhang Q Y, Zhu Q L, Liu W, Chen Y, Qiu R, Wang B, Yang Z F, Li H Y, Lin Y R, Xie Y Y, Shen R X, Chen S F, Wang Z, Chen Y L, Guo J X, Chen L T, Zhao X C, Dong Z C, Liu Y G. A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. Molecular Plant. 2015, 8(8): 1274-1284.