



# Genome-wide association study reveals early seedling vigour-associated quantitative trait loci in *indica* rice

Maosheng Zeng · Jing Yang · Kangjing Wu · Hao Wang · Kai Sun · Zhiqiang Chen · Tao Guo · Chun Chen

Received: 10 January 2021 / Accepted: 25 May 2021 / Published online: 11 June 2021  
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

**Abstract** Direct seeding is becoming a popular method of rice production in many rice-growing areas. Therefore, rice early seedling vigour (ESV) has also become an important agricultural research topic. To explore more genetic loci related to ESV, we used a germplasm resource containing 200 *indica* varieties genotyped by sequencing (GBS), and 161,657 high-quality SNPs were obtained. Nine phenotypes associated with ESV were evaluated. A genome-wide association study (GWAS) based on a mixed linear model (MLM) revealed 224 genetic loci for all traits, among which 70 genetic loci were repeatedly detected across two seasons and were reliable. In total, 1609 genes were predicted in the 200-kb genomic region harbouring these 70 loci. Furthermore, 16,184 differentially expressed genes

(DEGs) were identified using RNA sequencing (RNA-seq). With the combination of RNA sequencing and GWAS, 666 candidate DEGs were obtained. Based on gene annotation and quantitative real-time PCR (qRT-PCR), we predicted the 5 most promising candidate genes (*Os09g0278000*, *Os09g0279100*, *Os09g0279400*, *Os09g0279500* and *Os09g0280300*) at locus 167, which had a strong association signal. These findings provide valuable information for understanding the genetic control of ESV and for molecular marker-assisted breeding of *indica* rice.

**Keywords** *Indica* rice · GWAS · Early seedling vigour · RNA sequencing · Candidate genes

---

Maosheng Zeng, Jing Yang and Kangjing Wu have contributed equally to this work.

---

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10681-021-02868-y>.

---

M. Zeng · J. Yang · K. Wu · H. Wang ·  
K. Sun · Z. Chen · T. Guo (✉) · C. Chen (✉)  
National Engineering Research Center of Plant Space  
Breeding, South China Agricultural University, 483  
Wushan Road, Guangzhou, Guangzhou 510642, China  
e-mail: guo.tao@vip.163.com

C. Chen  
e-mail: chchun@scau.edu.cn

## Introduction

Seedling vigour refers to the ability of seeds to germinate rapidly from soil or water, mainly referring to the seed germination rate and early seedling growth and development (Huang et al. 2004). In recent years, rice direct seeding technology has been widely used in countries around the world. Compared with traditional rice cultivation technology, the new direct seeding cultivation technology has the advantages of labour savings and low costs, but there are still serious problems, including seedling death and poor crop establishment (Wu et al. 2006). As a result, ESV under

different cultivation techniques and related genetic analyses have attracted increasing attention.

Seedling vigour is an extremely complex agronomic trait that includes multiple indicators, such as the seed germination rate, germination potential and germination index (Wang et al. 2010), as well as the root length, seedling length, and wet and dry weights at the early seedling stage (Redoña and Mackill 1996; Regan et al. 1992). In summary, ESV is a quantitative trait controlled by multiple genes and is affected by various complex environmental factors. Quantitative trait locus (QTL) mapping has been proven to be an effective method to study complex traits (Huang et al. 2009; Song et al. 2007; Wang et al. 2015b). In recent years, many QTLs associated with seedling vigour have been detected. Wang et al. (2010) conducted QTL mapping analysis for the seed germination rate, final germination rate and germination index using a recombinant inbred population and obtained 10 QTLs accounting for 7.5% to 68.5% of the variation in the population. In the study of Dang et al. (2014), a total of 27 QTLs related to seed vigour were identified, and 15 hybrid combinations of excellent parents were configured to improve the seedling vitality. Xie et al. (2014) identified 8 QTLs associated with seedling vigour using an RIL population and further narrowed the associated QTLs down to two major QTLs, namely, *qSV-1* and *qSV-5c*, from 1.13 Mb and 400 kb, respectively. Abe et al. (2012) identified a QTL controlling seedling height at the long arm end of chromosome 3 and predicted that the candidate gene was *OsGA20ox1*. Compared with classical linkage analysis (QTL mapping), the number of SNP markers analysed by GWAS is huge (up to millions). Moreover, due to the abundant natural variation of the populations used for GWAS, GWAS can greatly shorten the time required for gene location, reduce the labour intensity, and effectively improve the efficiency of gene localization and cloning (Cordell and Clayton 2005; Zhang et al. 2010).

With the development of sequencing technology, an increasing number of genetic studies have been conducted on the quantitative traits of different species using high-density genetic maps (Poland et al. 2012; Sonah et al. 2013; Wang et al. 2015a; Xie et al. 2010). For example, Huang et al. (2010) reported that GWAS was used to study 14 complex agronomic traits related to rice morphology, yield, quality, seed colour and physiology using a population of 373 *indica* rice

germplasms and approximately 700,000 SNP genotypes. Cornell University's Susan R. Mc Couch's team used a high-density haploid mapping GWAS method with a natural population of 413 rice germplasms and 44,000 high-quality SNPs in the rice genome to study rice composite traits and detected that the rice blast resistance gene *Pita* is closely related to resistance to American rice blast races (IB-49, IC-17 and IE-1 K) (Zhao et al. 2011). In recent years, GWAS have also been successfully applied to rice seeds. Magwa et al. (2016) and Lu et al. (2017) determined the molecular mechanisms underlying seed dormancy. Sales et al. (2017), Schläppi et al. (2017) and Wang et al. (2017) analysed the regulatory mechanism of rice germination in a low-temperature environment. Hsu and Tung (2015) and Zhang et al. (2017) studied the molecular mechanism of rice germination in an anaerobic environment.

Seedling vigour is an extremely complex quantitative trait regulated by multiple genes. In order to locate the associated genetic loci, the population selected for the genome-wide association study (GWAS) analysis should have rich genetic diversity in its genetic background. The cultivated rice in Asia can be divided into two subspecies, *indica* and *japonica*. There are obvious differences in geographical distribution, morphology, physiology and biochemistry, and DNA molecular level between them. However, a large number of studies on rice genetic diversity found that the genetic diversity of *indica* rice was greater than that of *indica* rice (Jiang et al. 2000; Zhu et al. 2002; Qi et al. 2006). Moreover, *indica* rice in different rice growing areas had frequent gene exchange, complex genetic composition and diversified genetic background. At the same time, our research group is located in South China's *indica* rice region, we are mainly engaged in *indica* rice breeding. Therefore, the phenotypic identification of *indica* rice can quickly and directly serve our breeding work. In this study, phenotypic data from 200 germplasm resources collected worldwide were determined under optimal conditions (30 °C), and GWAS was performed in combination with genome-wide SNP marker analysis to identify reliable genetic loci affecting ESV.

## Materials and methods

### Plant materials

In this study, we selected 200 *indica* rice germplasm resources (Guo et al. 2019) from China (mainly from Guangdong Province), the International Rice Research Institute (IRRI), Vietnam, India, Nepal, South Korea and other countries and regions. All materials were planted in an experimental field at South China Agricultural University (located in Guangzhou, China, 23.16° N, 113.36° E). Twenty-five-day-old seedlings were transplanted to the experimental field. The plant spacing was 20 cm, and the row spacing was 20 cm. We selected plants of the same heading stage and estimated the heading stage of different varieties of rice when the main ear emerged from the leaf sheath (Jing et al. 2018). Considering the effect of seed ripening, we harvested 6 rice plants in the middle of each plot independently on the 35th day of the wet season (WS) after heading and the 40th day of the dry season (DS) after heading. The harvested seeds were dried for 5 days in a hot-air dryer at 42 °C and then stored at – 20 °C.

### Phenotypic evaluation of rice ESV

All methods were as described by Yang et al. (2019b). We collected WS and DS seeds of 200 *indica* rice germplasms and put them into an oven at 50 °C for 7 days of drying and heat treatment to break dormancy. The surfaces of all seeds were sterilized with 20% bleach (6–7% NaClO) for 20 min and then rinsed three times with sterile distilled water. In total, 400 seeds from the 200 WS and DS *indica* rice varieties were selected, and 50 seeds were selected from each group and placed in a petri dish covered with two layers of 9-cm-diameter circular filter paper. Ten millilitres of sterile distilled water was added, and the dishes were placed in incubators under a 12 h light/12 h dark cycle at a temperature of 30 °C. Approximately 2 days later, 8 seeds (the radicle or germ length of which reached approximately 1 mm) per plant were selected, placed in a germination box (length of 19 cm, width of 13 cm, height of 12 cm) and covered with two layers of filter paper, after which 20 ml of sterilized distilled water was added. The culture conditions were the same as those listed above. After 6 days, a WinRHIZO (Regent Instruments Inc.,

Québec, Canada) root image analysis system was used to measure the seedling stem diameter (SSD), seedling height (SH), root length (RL), root surface area (RSA), and root volume (RV). The shoot fresh weight (SFW), shoot dry weight (SDW), root fresh weight (RFW) and root dry weight (RDW) were measured by a sensitive balance. These nine indicators were used to evaluate the vigour of the rice seedlings. All data analysis was conducted using Statistical Analysis System (SAS) software version 9.4 for Windows (SAS Institute Inc. 2012). The frequency distribution for each trait was determined in Microsoft Excel 2016.

### GWAS mapping

Association mapping was performed according to the method in our published article (Guo et al. 2019). Tree Best (see <http://treesoft.sourceforge.net/treebest.shtml>) was used to build a p-distance-based neighbour-joined (NJ) tree. MEGA 6.0 (<http://www.megasoftware.net/>) software was used to visualize the phylogenetic trees. Pairwise relative kinship coefficients were calculated in TASSEL (ver. 4.0) (Bradbury et al. 2007).

We used the MLM in Genome-Wide Efficient Mixed Model Association (GEMMA) software to calculate the correlation value (*P*-value) of each SNP locus and trait based on genome-wide SNPs. For MLM analysis, we used the following formula:

$$y = X\alpha + S\beta + K\mu + e$$

In this equation, *y* is the phenotype, *X* is the genotype, *S* is the structure matrix, and *K* is the relationship matrix. *Xα* and *Sβ* represent fixed effects, and *Kμ* and *e* represent random effects. The top three principal components (PCs) were used to construct the *S* matrix for population structure correction. A matrix of simple matching coefficients was used to construct the *K* matrix.

### Verification of RNA-Seq by qRT-PCR

A good seedling vigour variety (GW27, YZX) was selected from 200 *indica* populations for RNA-Seq. All methods were as described by Yang et al. (2019a). The total RNA of each sample was homogenized in liquid nitrogen using a mortar and pestle and then purified using the Plant Total RNA Purification Kit (ComWin Biotech Company) according to the

manufacturer's instructions. The construction and sequencing of the RNA-seq library followed a previous protocol (Chen et al. 2016). Quality control was performed with Illumina HiSeq software, and all readings that passed the filter specifications were mapped to the reference genome IRGSP-1.0. After calculating the expression level of each transcript and gene, edgeR was used for differential expression analysis (Robinson et al. 2010). The FDR was used to determine the p-value threshold in multiple trials and to evaluate the significance of gene expression with an FDR threshold  $\leq 0.05$  and fold change absolute value  $\geq 2.0$ . RNA samples were reverse transcribed to cDNA using a high-capacity cDNA archive kit (Applied Biosystems, USA). The AceQ qPCR SYBR Green Master Mix Kit (Vazyme Biotech) was used to perform qRT-PCR according to the standard protocol, and the gene expression levels were determined on the StepOnePlus System (Applied Biosystems, USA). Three repetitions were used for each treatment. As an endogenous control, actin was used to normalize the obtained Ct values, and the relative expression values were calculated by the  $\Delta\Delta Ct$  method. The NCBI primer BLAST (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) was used to design gene-specific primers. The primers used for qRT-PCR are listed in Supplementary Table S1.

Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) enrichment analysis of DEGs

For the GO enrichment analysis of DEGs, we used the Singular Enrichment Analysis tool in AgriGO (Du et al. 2010) with the default parameters, and a  $P$ -value  $\leq 0.05$  indicated significant enrichment. The KEGG enrichment analysis of DEGs was performed using KOBAS software (Xie et al. 2011) with the default parameters, and a  $P$ -value  $\leq 0.05$  indicated significant pathway enrichment.

## Results

### Phenotypic variation of rice ESV

We calculated the mean value, maximum value, minimum value, standard deviation, kurtosis and skewness of the data for nine traits of 200 *indica* rice

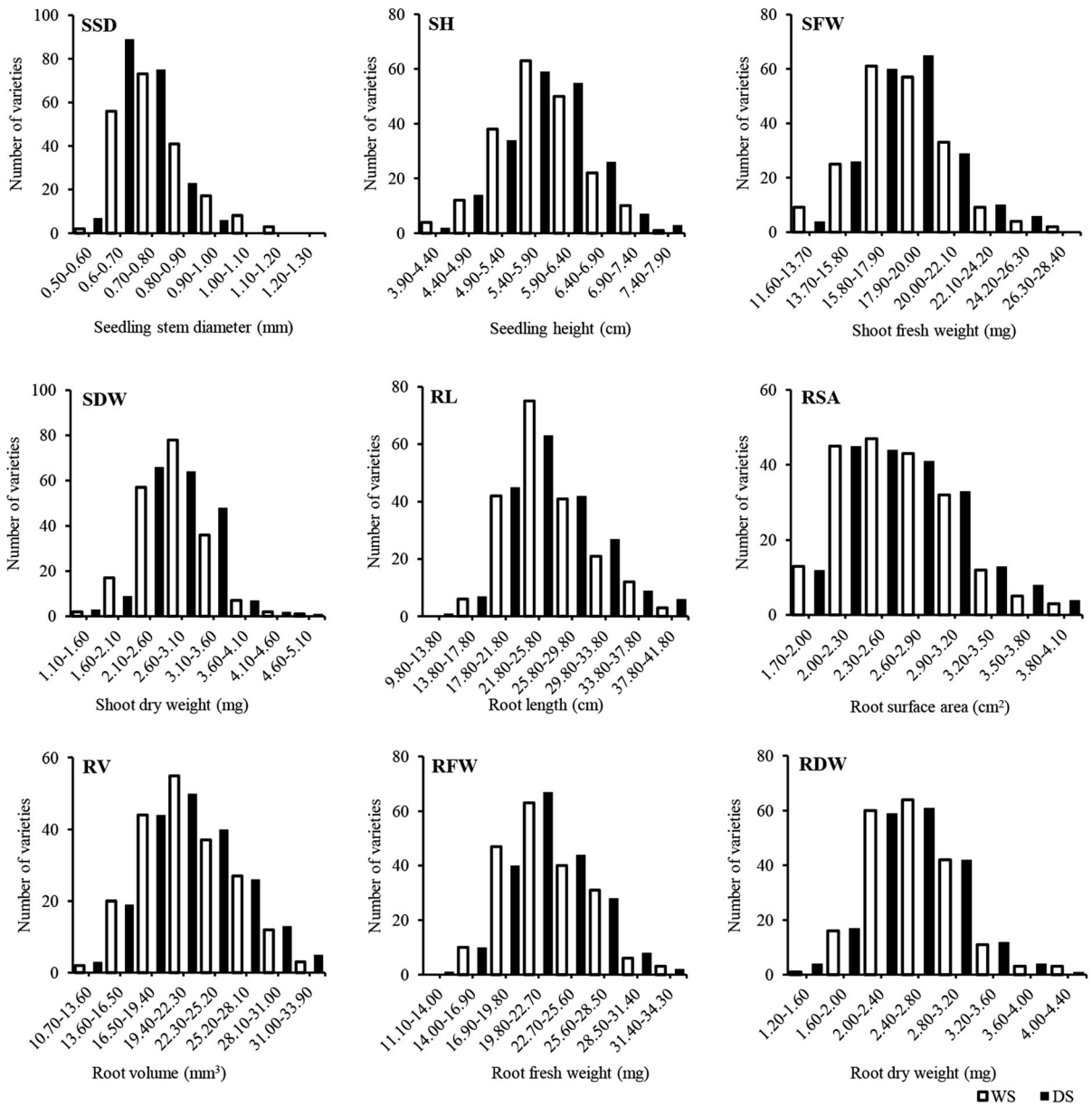
varieties (Fig. 1, Table S7) in the WS and DS. The results showed that the nine traits showed significant variation within this population (Table 1, Table S2). The coefficient of variation (CV) of seedling height in the WS was the smallest, at 11.12%. The CV of the root length in the DS was 21.21%. Except for the kurtosis and skewness of SSD in the WS and SFW in the WS and DS, which were slightly higher than 1.00, the absolute values of kurtosis and skewness of all traits were less than 1.00, indicating that all traits showed an approximately normal distribution. These observations showed that the traits related to ESV in this study were in line with the genetic law of quantitative traits.

### Correlation analysis of rice ESV traits.

To examine the relationships of the nine traits associated with ESV, we performed correlation analysis (Table 2). In the WS, the correlation coefficient between RV and RSA was the highest, at 0.859. In the DS, the correlation coefficient between RV and RSA was also the highest, at 0.828. Overall, the correlations among underground traits were higher than those among aboveground traits. We further observed the correlation coefficients among the four aboveground traits and five underground traits. The results showed that among the aboveground traits, consistent with our expectations, the correlation coefficient between SDW and SFW was the highest (WS, 0.770; DS, 0.687), and the correlation coefficient between SSD and SH was the lowest (WS, -0.196; DS, 0.002). Among underground traits, the correlation coefficient between RSA and RV was the highest (WS, 0.859; DS, 0.828). The correlation coefficient between RL and RV was the lowest in the WS, at 0.497, and the correlation coefficient between RL and RDW was the lowest in the DS, at 0.440.

### Population structure

In our previous study (unpublished), we sequenced the genomes of samples from 200 rice varieties collected around the world using GBS. The sequenced data were mapped to the Nipponbare reference genome (IRGSP 1.0) by BWA software. The average sequencing depth of the genome was 14.16X, and the coverage of at least one base was 11.24%. After filtering the data for a coverage depth  $\geq 2$ , minor allele frequency  $\geq 0.01$ ,



**Fig. 1** Histogram of the frequency distribution of SSD, SH, SFW, SDW, RL, RSA, RV, RFW and RDW in early rice seedlings

and missing rate  $\leq 0.9$ , we obtained a total of 161,657 high-quality SNPs.

First, the phylogenetic relationships of the 200 varieties were determined using the genetic distances calculated by SNPs. The resulting NJ tree (Fig. 2a) showed two branches (black and red) representing two subpopulations. GW102, GW178, GW189, GW143, GW169, and GW193 belonged to one subgroup, and the other subgroups contained 194 other varieties. Overall,

there was no significant group stratification among the 200 varieties. Furthermore, according to our calculations, most varieties had no meaningful relationship, and up to 99.91% of varieties possessed kinship values smaller than 0.10 (Fig. 2b), which indicated that the relative kinship of our population may not have had a strong influence on the GWAS. The analysis of the genetic structure and relative kinship made it clear that our *indica* population qualified for GWAS.

**Table 1** Mean (standard deviation), minimum, maximum, skewness and kurtosis of rice seedling SSD, SH, SFW, SDW, RL, RSA, RV, RFW, RDW

Trait <sup>a</sup>	Environment <sup>b</sup>	Mean	Range	Skewness	Kurtosis	CV <sup>c</sup> (%)
SSD (mm)	WS	0.77 ± 0.11	0.60–1.19	1.08	1.09	14.61
	DS	0.72 ± 0.08	0.55–1.00	0.94	0.89	11.33
SH (cm)	WS	5.77 ± 0.66	4.11–7.43	0.02	0.01	11.38
	DS	5.84 ± 0.65	3.98–7.71	0.10	0.24	11.12
SFW (mg)	WS	18.28 ± 2.73	11.62–27.74	0.54	0.70	14.92
	DS	18.41 ± 2.56	11.9–25.96	0.40	0.17	13.92
SDW (mg)	WS	2.76 ± 0.5	1.13–4.65	0.33	1.19	18.11
	DS	2.79 ± 0.5	1.43–4.67	0.38	1.13	18.06
RL (cm)	WS	25.16 ± 4.9	16.11–40.42	0.69	0.18	19.50
	DS	25.46 ± 5.4	9.88–41.18	0.65	0.45	21.21
RSA (cm <sup>2</sup> )	WS	2.61 ± 0.45	1.73–3.90	0.51	– 0.17	17.15
	DS	2.64 ± 0.48	1.73–4.08	0.60	– 0.01	17.97
RV (mm <sup>3</sup> )	WS	21.62 ± 4.14	13.37–33.27	0.41	– 0.37	19.15
	DS	21.72 ± 4.2	10.75–33.75	0.34	– 0.14	19.35
RFW (mg)	WS	22.15 ± 3.65	14.13–33.33	0.56	0.17	16.48
	DS	22.2 ± 3.67	11.18–33.56	0.38	0.21	16.54
RDW (mg)	WS	2.58 ± 0.46	1.47–4.19	0.68	0.81	18.03
	DS	2.56 ± 0.49	1.30–4.02	0.37	0.34	18.94

<sup>a</sup>Trait: SSD shoot stem diameter, SH shoot height, SFW shoot fresh weight, SDW shoot dry weight, RL root length, RSA root surface area, RV root volume, RFW root fresh weight, RDW root dry weight

<sup>b</sup>Environment: WS is the wet season in 2017; DS is the dry season in 2017

<sup>c</sup>CV coefficient of variation

**Table 2** Correlation analysis of rice seedling SSD, SH, SFW, SDW, RL, RSA, RV, RFW and RDW

Trait	SSD (mm)	SH (cm)	SFW (cm)	SDW (mg)	RL (cm)	RSA (cm <sup>2</sup> )	RV (mm <sup>3</sup> )	RFW (mg)	RDW (mg)
WS SSD (mm)		0.002	0.582**	0.345**	0.360**	0.496**	0.432**	0.519**	0.376**
SH (cm)	– 0.196**		0.574**	0.464**	0.253**	0.338**	0.292**	0.173*	0.215**
SFW (mg)	0.339**	0.595**		0.687**	0.392**	0.596**	0.552**	0.586**	0.468**
SDW (mg)	0.205**	0.487**	0.770**		0.266**	0.462**	0.470**	0.440**	0.387**
RL (cm)	0.218**	0.202**	0.354**	0.283**		0.827**	0.528**	0.545**	0.440**
RSA (cm <sup>2</sup> )	0.316**	0.297**	0.582**	0.495**	0.838**		0.828**	0.700**	0.581**
RV (mm <sup>3</sup> )	0.295**	0.282**	0.635**	0.550**	0.497**	0.859**		0.608**	0.554**
RFW (mg)	0.398**	0.205**	0.610**	0.467**	0.582**	0.767**	0.719**		0.718**
RDW (mg)	0.283**	0.197**	0.486**	0.467**	0.489**	0.654**	0.604**	0.694**	

\*, \*\*Significant at the 0.05 and 0.01 levels, respectively

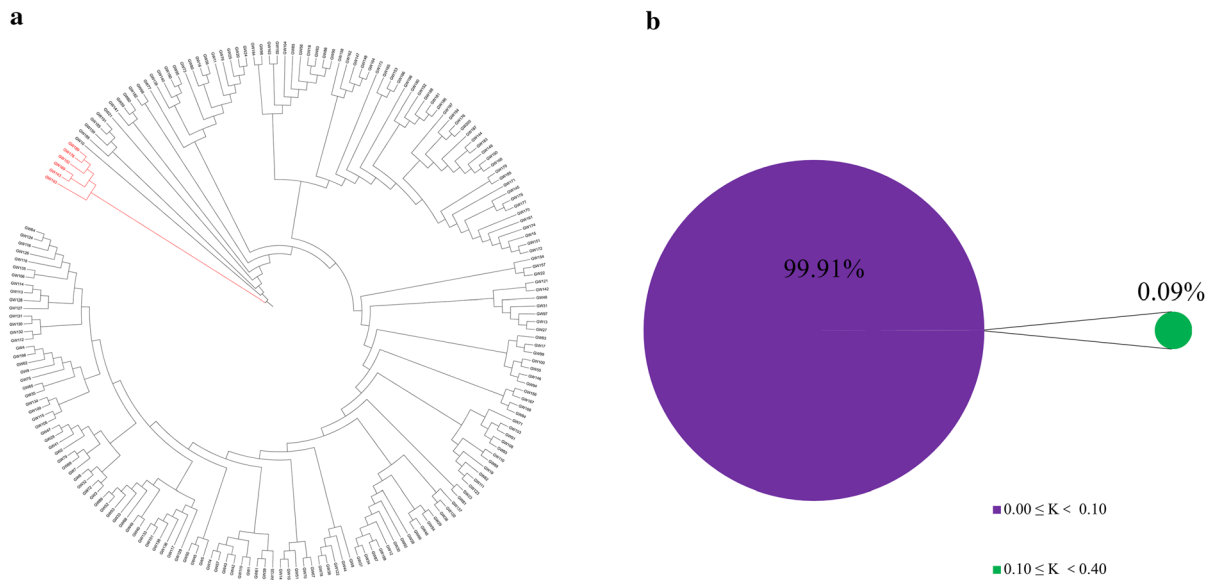
WS wet season, DS dry season

Genome-wide association analysis of ESV

To determine the genetic loci associated with rice ESV, we used an MLM for GWAS of SSD, SH, SFW, SDW, RL, RSA, RV, RFW and RDW to obtain a

Manhattan diagram and quantile–quantile diagram (Fig. 3 and Fig. 4).

We obtained a total of 345 significant SNPs located across all 12 chromosomes, indicating that the vigour of rice seedlings is a very complex quantitative trait. Among them, the number of SNPs detected by SSD



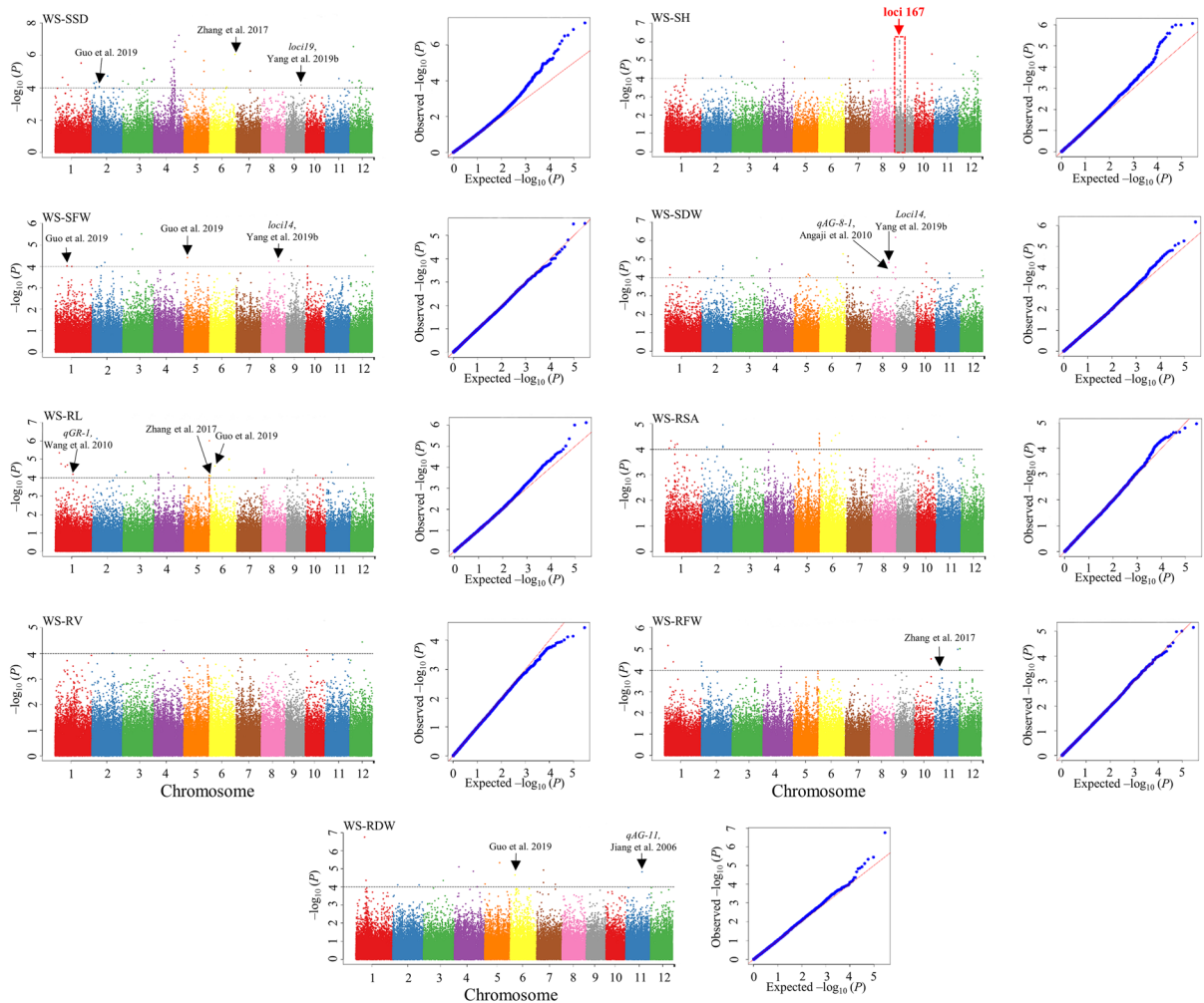
**Fig. 2** Population structure of 200 rice cultivars. **a** NJ tree based on the p-distance. **b** Distribution of relative kinship among 200 *indica* accessions, where K represents the relative kinship coefficients

was the highest, at 76, and the number of SNPs detected by RV was the lowest, at only 16 (Fig. 5a). We analysed the traits associated with the detected SNPs and found that these SNPs were mainly associated with RL, SDW, RSA, SSD and SH. We mapped these SNPs onto each chromosome and found that the SNPs of these five traits accounted for a large proportion of each chromosome (Fig. 5b). We grouped SNPs with the same physical location into the same genetic loci and finally detected 224 genetic loci (Supplementary Table S3). The distribution of these genetic loci on chromosomes was further analysed. We found that loci were detected on all chromosomes, of which chromosome 4 had the most detections, with a total of 31, and chromosome 7 had the least, with a total of 10. (Fig. 5c). Remarkably, 70 genetic loci were repeated over the two seasons (Fig. 5d, Table 3).

#### Comparison of GWAS results with previously reported loci

To verify the accuracy of our results, we compared the significant SNP sites in this study with those from previous studies. In our study (Supplementary Table S3 and Fig. 6), in total, 23 SNPs were located in the same areas as those highlighted in previous studies, among which 14 genetic loci were identified.

Locus 13, which is associated with SDW and SFW, was identified in the genomic interval of S1–13,783,629 (Guo et al. 2019), and the physical distance was very close, at only 19 bp. Locus 16, which is associated with RL, was identified in the genomic interval of *qGR-1*, which is related to the GR at 3 days (Wang et al. 2010). Locus 39, which is associated with SSD, was identified in the genomic interval S2-8,475,224 (Guo et al. 2019), and the physical distance between the two sites was 118 kb. Locus 49, which is associated with SSD, was identified in the genomic interval of *qSDW2* (Han et al. 2007) and Locus 4 (Yang et al. 2019b). Locus 109, which is associated with SFW and RSA, was identified in the genomic interval of S5-3,878,312, which was related to VI (Guo et al. 2019). Locus 125, which was associated with the RL, was identified in the genomic interval of seq-rs2683, which was related to the FTI, and the physical distance between the two sites was 26 kb (Zhang et al. 2017). Locus 126, which was associated with RL and RDW, was identified in the genomic interval S6-5,542,656 (Guo et al. 2019), and the physical distance between the two sites was 77 kb. Locus 139, which was associated with SSD, was identified in the genomic interval of seq-rs3121 (Zhang et al. 2017). Locus 150, which was associated with SSD, was identified in the genomic interval of *qAG-7-2* (Angaji et al. 2010) and id7005515 (Hsu and



**Fig. 3** Genome-wide association mapping of SSD, SH, SFW, SDW, RL, RSA, RV, RFW, and RDW of wet-season rice seedlings was performed by the MLM method. The horizontal dashed line

represents the significance threshold  $-\log_{10}(P) = 4.00$ . The black text indicates the published QTLs and SNPs, and the red arrow represents locus 167, which had a strong association signal

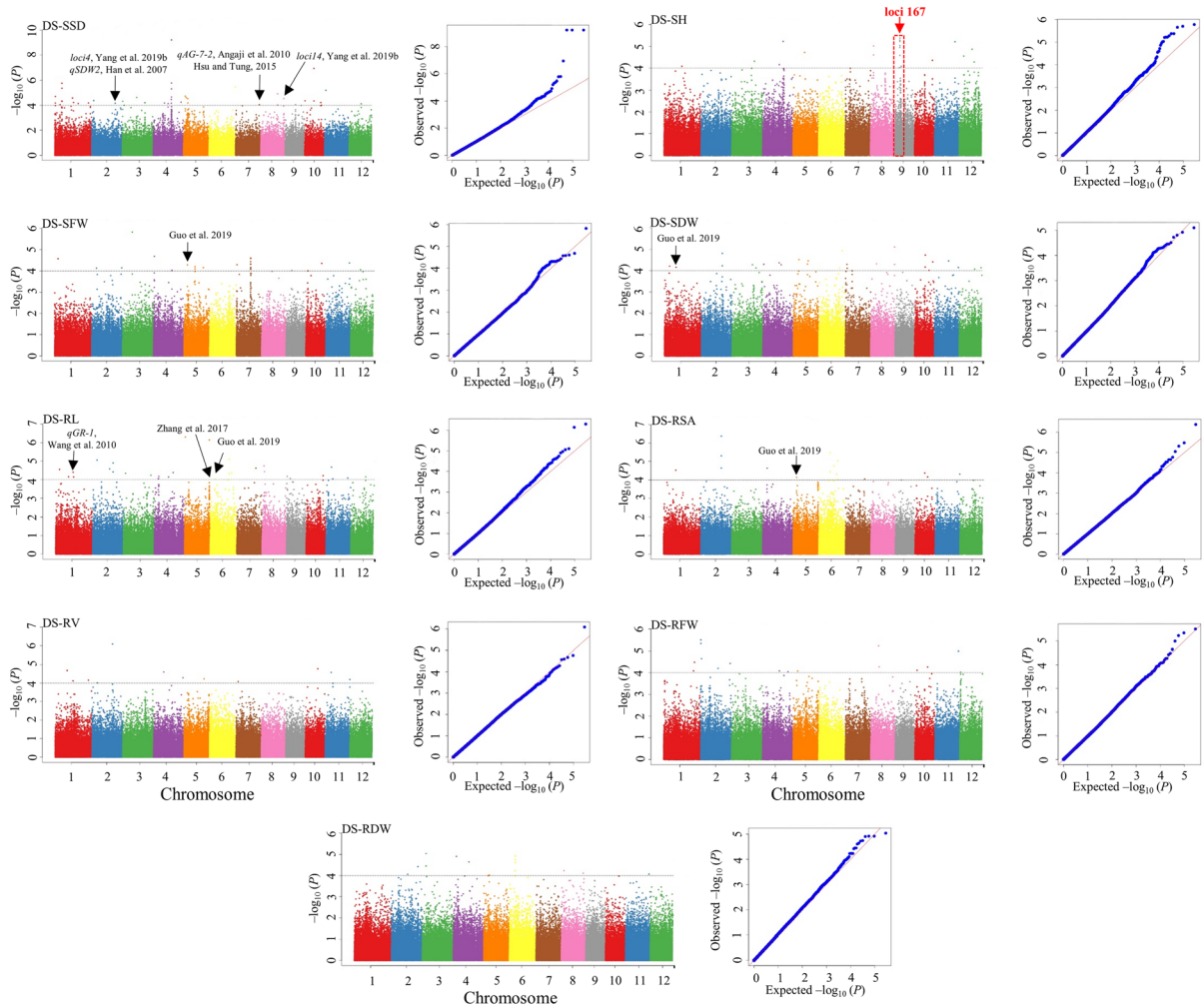
Tung, 2015). Locus 159, which was associated with SSD, SDW, and SFW, was identified in the genomic interval of Locus 14 (Yang et al. 2019b). Locus 160, which was associated with SDW, was identified in the genomic interval of *qAG-8-1* (Angaji et al. 2010), and the physical distance between the two sites was 55 kb. Locus 175, which was associated with SSD, was identified in the genomic interval of Locus 19 which was related to RL (Yang et al. 2019b). Locus 193, which was associated with RFW, was identified in the genomic interval of seq-rs4859, which was related to the FTI, and the physical distance between the two sites was 35 kb (Zhang et al. 2017). Locus 198, which

was associated with RDW, was identified in the genomic interval of *qAG-11* (Jiang et al. 2006), and the physical distance between the two sites was 53 kb. The above results further prove the reliability of our research.

#### Identification of candidate genes

To further narrow down the candidate gene screening range, we further divided SNPs with the same physical location into the same genetic locus and finally obtained 224 genetic loci (Table S3). Of these 224 genetic loci, 70 were located in both the WS and DS,



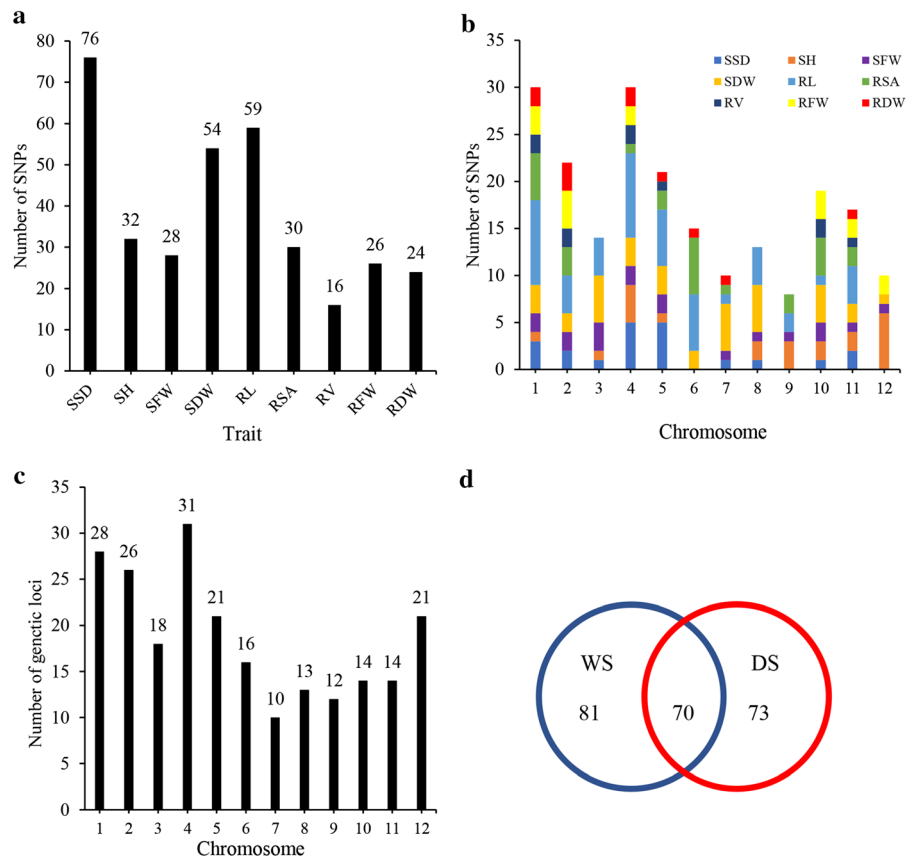


**Fig. 4** Genome-wide association mapping of SSD, SH, SFW, SDW, RL, RSA, RV, RFW, and RDW of dry-season rice seedlings was performed by the MLM method. The horizontal dashed line

represents the significance threshold  $-\log_{10}(P) = 4.00$ . The black text indicates the published QTLs and SNPs, and the red arrow represents locus 167, which had a strong association signal

making them reliable genetic loci. From these 70 loci, we screened the candidate genes that affect the vigour of rice seedlings. Rice is a self-intersecting plant, so it has a certain decay rate. The LD value of rice is limited to  $\sim 200$  kb for mapping resolution (Huang et al. 2012). We targeted the regions 200 kb upstream and downstream of each SNP. The interval of these 70 loci contained a total of 1,609 genes. Moreover, we used the variety GW27 with a better ESV in this population and conducted RNA-seq analysis on seeds germinating on day 0 and day 4 at room temperature (30 °C) (Guo et al. 2019). By analysing the differential expression of the transcription spectra of GW27 at 0

and 4 d (absolute value of a fold-change  $\geq 2$ ,  $FDR \leq 0.05$ ), a total of 16,184 DEGs was obtained, among which 12,194 were upregulated, accounting for 75.35% of the total DEGs, and 3,990 were downregulated, accounting for 24.65% of the total DEGs (Fig. 7a). Combined with GWAS and RNA-seq analysis, in total, 666 overlapping DEGs were screened (Fig. 7b and Supplementary Table S4).



**Fig. 5** Summary of 345 SNPs associated with 9 traits by GWAS. **a** Number of SNPs associated with traits. **b** Number of SNPs detected on chromosomes. **c** Number of genetic sites on

chromosomes. **d** In total, 70 overlapping DEGs were identified by both RNA sequencing and GWAS

KEGG and GO enrichment analysis of the 666 DEGs

We applied GO to classify the functions of the 666 DEGs. For the identification of metabolic pathways in which the 666 DEGs were involved and enriched, the KEGG database was used.

GO enrichment analysis of DEGs was conducted with an  $FDR \leq 0.05$  as the threshold. These DEGs were mainly concentrated in the molecular function, biological process and cellular component categories (Table 3). In the biological processes (BP) category, 2 pathways were significantly enriched, and the DEGs were mainly involved in processes such as the cellular response to chemical stimulus (GO:0,070,887) and intracellular signalling cascade (GO:0,007,242). In the molecular function (MF) category, 6 pathways were significantly enriched, and the DEGs were mainly related to receptor activity (GO:0,004,872), molecular

transducer activity (GO:0,060,089), and signal transducer activity (GO:0,004,871). In the cellular component (CC) category, 12 pathways were significantly enriched, and the DEGs were mainly in the cytoplasmic part (GO:0,044,444), cytoplasm (GO:0,005,737), intracellular membrane-bound organelle (GO:0,043,231), and membrane-bound organelle (GO:0,043,227) categories.

We set an  $FDR \leq 0.05$  as the threshold for screening DEGs, but this did not yield any significantly enriched pathways. Therefore, we used a  $P$ -value  $\leq 0.05$  as the threshold for screening DEGs in the analysis and obtained the following results (Supplementary Fig. S1). We found that these DEGs were annotated into 9 signalling pathways, which mainly included cyanoamino acid metabolism, diterpenoid biosynthesis, starch and sucrose metabolism, circadian rhythm—plant, phenylpropanoid biosynthesis, and DNA replication.

**Table 3** GO enrichment analysis of the 666 differentially expressed genes (DEGs)

Ontology	GO term	Description	P-value	FDR
BP	GO:0070887	Cellular response to chemical stimulus	1.30E-06	7.80E-04
	GO:0007242	Intracellular signalling cascade	1.50E-04	4.40E-02
MF	GO:0004872	Receptor activity	5.80E-11	2.20E-08
	GO:0060089	Molecular transducer activity	1.80E-09	2.30E-07
	GO:0004871	Signal transducer activity	1.80E-09	2.30E-07
	GO:0004713	Protein tyrosine kinase activity	1.60E-06	1.50E-04
	GO:0016859	Cis–trans isomerase activity	5.40E-04	3.50E-02
	GO:0003755	Peptidyl–prolyl cis–trans isomerase activity	5.40E-04	3.50E-02
CC	GO:0044444	Cytoplasmic part	1.30E-260	1.40E-258
	GO:0005737	Cytoplasm	1.60E-236	8.60E-235
	GO:0043231	Intracellular membrane–bounded organelle	1.30E-202	4.70E-201
	GO:0043227	Membrane–bounded organelle	2.40E-201	6.50E-200
	GO:0043229	Intracellular organelle	2.60E-178	4.80E-177
	GO:0043226	Organelle	2.60E-178	4.80E-177
	GO:0044424	Intracellular part	3.60E-161	5.70E-160
	GO:0005622	Intracellular	2.00E-150	2.70E-149
	GO:0044464	Cell part	3.50E-142	3.90E-141
	GO:0005623	Cell	3.50E-142	3.90E-141
	GO:0005739	Mitochondrion	1.10E-72	1.10E-71
GO:0016020	Membrane	1.90E-05	1.70E-04	

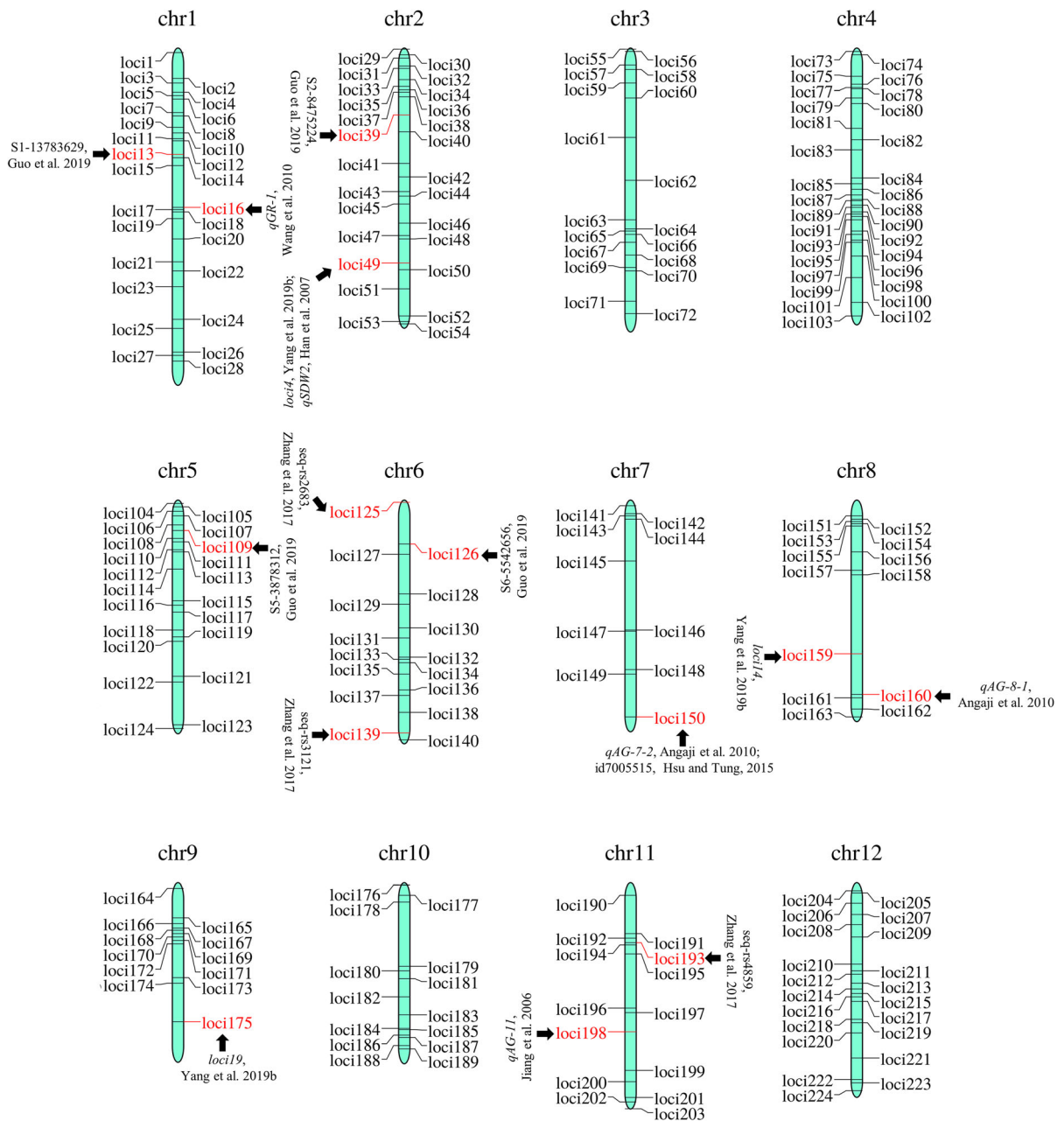
BP biological processes, MF molecular function, CC cellular component

We further focused on locus 167 on chromosome 9, which was associated with SH, appeared steadily in both the WS and DS and had a strong association signal ( $-\log P > 5$ ). There were only 16 genes in this region, as further based on the transcriptome expression profile, and 5 differentially expressed genes. We think that these 5 DEGs are the most promising candidates, including the seed maturation protein LEA 4 (*Os09g0278000*), AP2-associated protein kinase 1 (*Os09g0279100*), rhodanese-like (*Os09g0279400*), RNA recognition motif containing protein (*Os09g0279500*), and oxidation domain protein (*Os09g0280300*). Using the differences in the qRT-PCR results from YZX (Yuzhenxiang) germinated seeds from the fourth day of the experiment, we further verified the expression levels of these five genes (Fig. 8 and Supplementary Table S5). The qRT-PCR results were similar to the RNA-seq analysis results, indicating that our RNA-seq results were reliable. In summary, our results showed that through the combination of RNA-seq analysis, GWAS, GO

enrichment analysis, and KEGG enrichment analysis, we identified the five most likely candidate genes related to the vigour of rice seedlings.

## Discussion

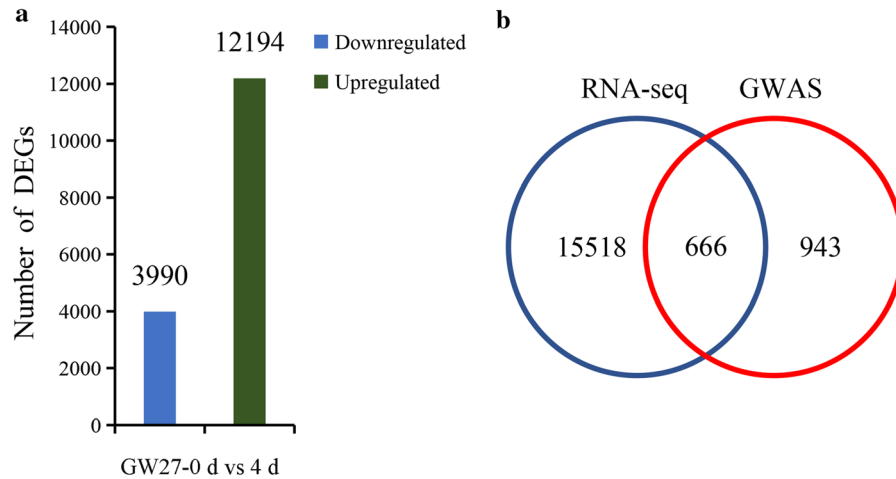
With the development of society, the economy, and technology, labour and production costs are increasing constantly. Direct rice seeding is becoming increasingly popular around the world as a light and simplified cultivation mode. For a long time, breeders have been breeding for the characteristics of transplanted rice, so many varieties have difficulty adapting to the requirements of direct-seeding rice cultivation. In particular, problems such as difficult germination of seeds in the field, serious weeds, and lodging have severely restricted the promotion of direct-seeding rice (Zhu and Yan 1997). Therefore, it is of great significance to identify new genetic loci related to seedling vigour for direct-seeding rice breeding.



**Fig. 6** Loci collocated with those in previous studies

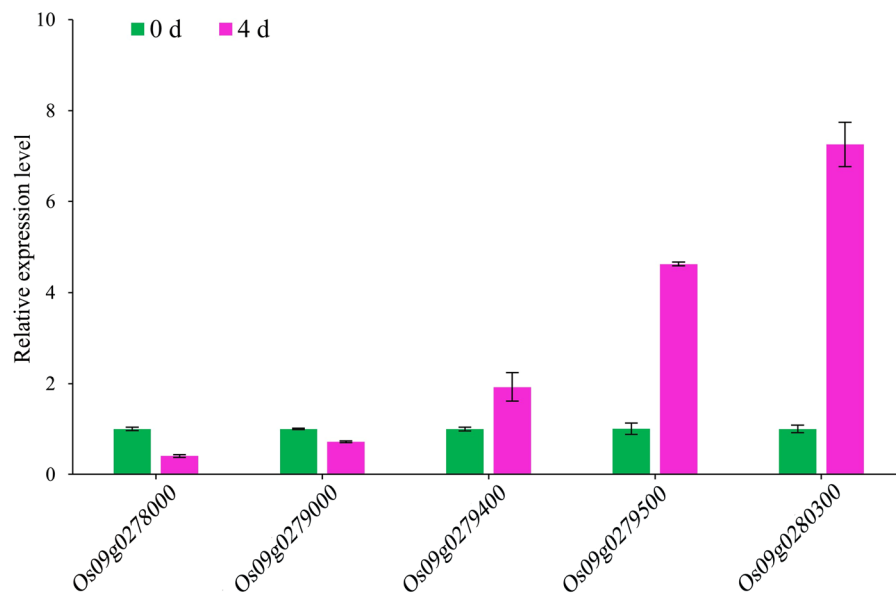
Six-day-old rice seedlings are in a critical period during which rice with strong ESV can successfully compete with weeds for nutrition and lay a solid foundation for high and stable yields in the later stage. Here, we measured a total of nine traits, SSD, SH, RL, RSA, RV, SFW, SDW, RFW, and RDW, after seedlings grew for 6 days. These 9 indexes can

effectively and comprehensively evaluate the ESV. In addition, we used a total of 200 materials for the GWAS. According to the phenotypic statistics of the nine rice seedling traits examined in this study, the phenotypes of various traits were abundant, and the differences between the materials were large, indicating that these are typical quantitative traits (Table 1),



**Fig. 7** Candidate gene identification by integrating GWAS and RNA sequencing analyses. **a** The number of differentially expressed genes (DEGs) detected by RNA sequencing on day 4 of germination. **b** Genome-wide association (GWAS) combined

with RNA-seq analysis was used to screen candidate genes. A total of 666 overlapping DEGs were identified by RNA-seq analysis and GWAS



**Fig. 8** Expression profiles of the five most likely candidate genes detected by qRT-PCR. 0 d, imbibition starting point. 4 d, the 4th day of germination

and the materials within the population may contain superior alleles.

Compared with linkage analysis, association analysis uses the natural population, which is simple to construct, rich in genetic diversity, and high in detection efficiency. Generally, more genetic loci can be obtained than linkage analysis. In this study, using GBS to sequence 200 rice varieties, a total of

161,657 high-quality SNPs were obtained. GWAS was performed based on an MLM, and 224 genetic loci were obtained (Supplementary Table S3). The accuracy of GWAS is also high. For example, in the region of locus 167, there are only 16 genes, which is conducive to the next step of fine mapping.

The transcriptome refers to the sum of the RNA transcribed by cells or tissues of an organism under

specific space–time conditions (Zhang et al. 2012). As far as the genome is concerned, the transcriptome, which can be used to study the functional genes and their metabolic pathways more effectively, is only concerned with the genes being transcribed (Zhou et al. 2012). With the development of high-throughput sequencing technology in recent years, the sequencing time and cost have been greatly reduced (Shendure and Ji 2008). In this study, 12,194 DEGs with upregulated expression and 3990 DEGs with downregulated expression were obtained through transcriptome sequencing. We detected 666 differentially expressed genes in the 70 locus repeats at WS and DS, with an average of less than 10 genes at each locus. In particular, locus 167, upon which we focused, had only 5 DEGs among 16 genes in this region. This greatly reduces the range of candidate genes, and we can use reverse genetics to knock out these genes and obtain the corresponding mutants to identify the target genes.

Based on quantitative qRT-PCR, we predicted 5 candidate genes (*Os09g0278000*, *Os09g0279100*, *Os09g0279400*, *Os09g0279500* and *Os09g0280300*) at locus 167. The seed maturation protein LEA 4, which is encoded by the late embryogenesis abundant protein gene (*Os09g0278000*), is expressed during seed maturation and development (Lu et al. 2013). It is responsible for regulating the expression of related proteins during seed maturation. *Os09g0279400* (rhodanese-like) and *Os09g0280300* (oxidoreductase domain protein) showed similar expression patterns to previously reported expression profiles obtained from seeds with aerobic germination treatment (Howell et al. 2008), i.e., the expression levels were upregulated rapidly during germination. Based on these results taken together, these five DEGs are the candidates most responsible for ESV.

## Conclusions

In this study, we used a natural population containing 200 *indica* rice materials. In total, 161,657 high-quality SNPs were obtained by the GBS method, and GWAS was conducted for nine traits related to ESV. In total, 224 genetic loci were obtained in the two seasons, of which 70 appeared repeatedly in the two seasons and were relatively stable. According to the gene expression profile, 666 differentially expressed

candidate genes were identified in the 70 stable intervals. Five candidate genes in locus 167 were obtained using qRT-PCR. These findings have laid the foundation for our molecular breeding of direct-seeding rice.

**Acknowledgements** The authors thank the Guangdong Rural Revitalization Project (yuecainong (2020) 39), the Research and Development Plan for Key Areas in Guangdong Province (No. 2018B020206002) and the National Key Technology Research and Development Program (No. 2016YFD102102) of the Ministry of Science and Technology of the People's Republic of China for their support.

**Authors' contributions** ZQC, TG and CC conceived the study. MSZ, JY and KJW performed all the experiments and wrote the manuscript. KS and HW assisted in conducting the experiments and analysing the data.

**Funding** This research was supported by the Research and Development Plan for Key Areas in Guangdong province (No. 2018B020206002), the Guangdong Rural Revitalization Project (yuecainong (2020) 39) and the National Key Technology Research and Development Program (No. 2016YFD102102) of Ministry of Science and Technology of the People's Republic of China. The funding agency had no input into the experimental design, the conduct of the research, the analysis and interpretation of experimental results or the writing of the manuscript.

**Data availability** All data generated or analysed during this study are included in this published article.

## Declarations

**Conflicts of interest** The authors declare that there are no competing interests.

**Ethics approval** Not applicable.

**Consent for publication** Not applicable.

**Grant Disclosures** The following grant information was disclosed by the authors: Guangdong Rural Revitalization Project: yuecainong (2020) 39. Research and Development Plan for Key Areas in Guangdong Province: No. 2018B020206002. National Key Technology Research and Development Program: No. 2016YFD102102.

## References

- Abe A, Takagi H, Fujibe T et al (2012) *OsGA20ox1*, a candidate gene for a major QTL controlling seedling vigor in rice. *Theor Appl Genet* 125(4):647–657. <https://doi.org/10.1007/s00122-012-1857-z>
- Angaji SA, Septiningsih EM, Mackill DJ et al (2010) QTLs associated with tolerance of flooding during germination in

- rice (*Oryza sativa* L.). *Euphytica* 172:159–168. <https://doi.org/10.1007/s10681-009-0014-5>
- Bradbury PJ, Zhang ZW, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* (Oxford, England) 23(19):2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Chen LK, Gao WW, Chen SP et al (2016) High-resolution QTL mapping for grain appearance traits and co-localization of chalkiness-associated differentially expressed candidate genes in rice. *Rice* (New York, N.Y.) 9:48. <https://doi.org/10.1186/s12284-016-0121-6>
- Cordell HJ, Clayton DG (2005) Genetic association studies. *The Lancet* 366(9491):1121–1131. [https://doi.org/10.1016/S0140-6736\(05\)67424-7](https://doi.org/10.1016/S0140-6736(05)67424-7)
- Dang XJ, Thi TGT, Dong GS et al (2014) Genetic diversity and association mapping of seed vigor in rice (*Oryza sativa* L.). *Planta* 239(6):1309–1319. <https://doi.org/10.1007/s00425-014-2060-z>
- Du Z, Zhou X, Ling Y, Zhang ZS (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucl Acids Res* 38(suppl\_2): W64–W70. <https://doi.org/10.1093/nar/gkq310>
- Guo T, Yang J, Li DX et al (2019) Integrating GWAS, QTL, mapping and RNA-seq to identify candidate genes for seed vigor in rice (*Oryza sativa* L.). *Mol Breed* 39:87. <https://doi.org/10.1007/s11032-019-0993-4>
- Han LZ, Qiao YL, Zhang SY et al (2007) Identification of quantitative trait loci for cold response of seedling vigor traits in rice. *J Genet Genomics* 34(3):239–246. [https://doi.org/10.1016/S1673-8527\(07\)60025-3](https://doi.org/10.1016/S1673-8527(07)60025-3)
- Howell KA, Narsai R, Carroll A et al (2008) Mapping metabolic and transcript temporal switches during germination in Rice highlights specific transcription factors and the role of RNA instability in the germination process. *Plant Physiol* 149:961–980. <https://doi.org/10.1104/pp.108.129874>
- Hsu SK, Tung CW (2015) Genetic mapping of anaerobic germination-associated QTLs controlling coleoptile elongation in rice. *Rice* (New York, N.Y.) 8(1):38. <https://doi.org/10.1186/s12284-015-0072-3>
- Huang Z, Yu T, Su L et al (2004) Identification of chromosome regions associated with seedling vigor in rice. *Yi Chuan Xue Bao* 31(6):596–603. <https://doi.org/10.1088/1009-0630/6/5/011>
- Huang XZ, Qian Q, Liu ZB et al (2009) Natural variation at the *DEP1* locus enhances grain yield in rice. *Nat Genet* 41(4):494–497. <https://doi.org/10.1038/ng.352>
- Huang XH, Wei XH, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42(11):961–967. <https://doi.org/10.1038/ng.695>
- Huang XH, Zhao Y, Wei XH et al (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* 44(1):32–39. <https://doi.org/10.1038/ng.1018>
- Jiang YB, Chen L, Wu CM et al (2000) Study on the relationship between heterosis and genetic differentiation in rice. *Acta Agronomica Sinica* 26(6):641–649
- Jiang L, Liu SJ, Hou MY et al (2006) Analysis of QTLs for seed low temperature germinability and anoxia germinability in rice (*Oryza sativa* L.). *Field Crops Res* 98(1):68–75. <https://doi.org/10.1016/j.fcr.2005.12.015>
- Jing L, Xu R, Wang CC et al (2018) A heading date QTL, *qHD7.2*, from wild rice (*Oryza rufipogon*) delays flowering and shortens panicle length under long-day conditions. *Sci Rep* 8(1):2928. <https://doi.org/10.1038/s41598-018-21330-z>
- Lu S, Yang N, Xiong TY (2013) Advances in researches on LEA proteins and lea genes. *J Sichuan Forestry Sci Technol.* <https://doi.org/10.16779/j.cnki.1003-5508.2013.03.005>
- Lu Q, Niu XJ, Zhang MC et al (2017) Genome-wide association study of seed dormancy and the genomic consequences of improvement footprints in rice (*Oryza sativa* L.). *Front Plant Sci* 8:2213. <https://doi.org/10.3389/fpls.2017.02213>
- Magwa RA, Zhao H, Xing YZ (2016) Genome-wide association mapping revealed a diverse genetic basis of seed dormancy across subpopulations in rice (*Oryza sativa* L.). *BMC Genetics* 17(1):28. <https://doi.org/10.1186/s12863-016-0340-2>
- Poland JA, Brown PJ, Sorrells ME et al (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7(2):e32253
- Qi YW, Zhang DL, Zhang HL et al (2006) Genetic diversity of selected rice varieties in China and its trend in recent 50 years. *Chin Sci Bull* 51(6):693–699
- Redoña ED, Mackill DJ (1996) Mapping quantitative trait loci for seedling vigor in rice using RFLPs. *Theor Appl Genet* 92(3–4):395–402. <https://doi.org/10.1007/BF00223685>
- Regan KL, Siddique KHM, Turner NC et al (1992) Potential for increasing early vigour and total biomass in spring wheat. II. Characteristics associated with early vigour. *Aus J Agric Res* 43(3):541–553. <https://doi.org/10.1071/ar9920541>
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (Oxford, England) 26(1):139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Sales E, Viruel J, Domingo C et al (2017) Genome wide association analysis of cold tolerance at germination in temperate japonica rice (*Oryza sativa* L) varieties. *PLoS One* 12(8): e183416 <https://doi.org/10.1371/journal.pone.0183416>
- Schläppi MR, Jackson AK, Eizenga GC et al (2017) Assessment of five chilling tolerance traits and GWAS mapping in rice using the USDA mini-core collection. *Front Plant Sci* 8:957. <https://doi.org/10.3389/fpls.2017.00957>
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26(10):1135–1145. <https://doi.org/10.1038/nbt1486>
- Sonah H, Bastien M, Iquiria E et al (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* 8(1): e54603. Doi: <https://doi.org/10.371/journal.pone.0054603>
- Song XJ, Huang W, Shi M et al (2007) A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat Genet* 39(5):623–630. <https://doi.org/10.1038/ng2014>
- Wang ZF, Wang JF, Bao YM et al (2010) Quantitative trait loci analysis for rice seed vigor during the germination stage. *J Zhejiang Univ Sci B* 11(12):958–964. <https://doi.org/10.1631/jzus.B1000238>

- Wang XD, Yu KJ, Li HG et al (2015a) High-density SNP map construction and QTL identification for the apetalous character in *brassica napus* L. *Front Plant Sci* 6:1164. <https://doi.org/10.3389/fpls.2015.01164>
- Wang YX, Xiong GS, Hu J et al (2015b) Copy number variation at the *GL7* locus contributes to grain size diversity in rice. *Nat Genet* 47(8):944–948. <https://doi.org/10.1038/ng.3346>
- Wang X, Zou BH, Shao QL et al (2017) Natural variation reveals that *OsSAP16* controls low-temperature germination in rice. *J Exp Bot* 69(3):413–421. <https://doi.org/10.1093/jxb/erx413>
- Wu WG, Chen Y, Qian YF et al (2006) The current status and progresses of the research on direct seeding rice. *Rev China Agric Sci Technol* 2006(04):32–36. <https://doi.org/10.3969/j.issn.1008-0864.2006.04.007>
- Xie WB, Feng Q, Yu HH et al (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci USA* 107(23):10578–10583. <https://doi.org/10.1073/pnas.1005931107>
- Xie C, Mao JJ, Huang Y, Ding JM et al (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucl Acids Res* 39(suppl\_2):W316–W322. <https://doi.org/10.1093/nar/gkr483>
- Xie LX, Tan ZW, Zhou Y et al (2014) Identification and fine mapping of quantitative trait loci for seed vigor in germination and seedling establishment in rice. *J Integr Plant Biol* 56(8):749–759. <https://doi.org/10.1111/jipb.12190>
- Yang J, Sun K, Li DX et al (2019a) Identification of stable QTLs and candidate genes involved in anaerobic germination tolerance in rice via high-density genetic mapping and RNA-Seq. *BMC Genomics* 20(1):355. <https://doi.org/10.1186/s12864-019-5741-y>
- Yang J, Yang GL, Yang M et al (2019b) Quantitative trait locus analysis of seed germination and early seedling growth in rice. *Front Plant Sci* 10:1582. <https://doi.org/10.3389/fpls.2019.01582>
- Zhang ZW, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42(4):355–360. <https://doi.org/10.1038/ng.546>
- Zhang CL, Qin ZJ, Wang GZ et al (2012) Transcriptome and RNA-seq technology. *Biotechnol Bulletin* 2012(12):51–56. <https://doi.org/10.13560/j.cnki.biotech.bull.1985.2012.12.025>
- Zhang MC, Lu Q, Wu W et al (2017) Association mapping reveals novel genetic loci contributing to flooding tolerance during germination in *indica* rice. *Front Plant Sci* 8:678. <https://doi.org/10.3389/fpls.2017.00678>
- Zhao KY, Tung CW, Eizenga GC et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* 2(1):467. <https://doi.org/10.1038/ncomms1467>
- Zhou H, Zhang X, Liu TY et al (2012) Data processing and gene discovery of high-throughput transcriptome sequencing. *Jiangxi Sci* 30(05):607–611. <https://doi.org/10.13990/j.issn1001-3679.2012.05.035>
- Zhu DF, Yan XQ (1997) Survey of rice direct sowing and cultivation in overseas. *Cultivat Cropping* 1997(Z1):102–103. <https://doi.org/10.13605/j.cnki.52-1065/s.1997>
- Zhu ZF, Sun CQ, Fu YC et al (2002) Comparison of genetic diversity between *Oryza sativa* L. and *Oryza rufipogon* Griff. by SSR markers. *Sci Agric Sin* 35(12):1437–1441

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