Molecular Genetics and Genomics
https://doi.org/10.1007/s00438-020-01704-7

ORIGINAL ARTICLE

Genome-wide association mapping for protein, oil and water-soluble protein contents in soybean

Shanshan Zhang[1] · Derong Hao[2] · Shuyu Zhang[1] · Dan Zhang[3] · Hui Wang[1] · Haiping Du[4] · Guizhen Kan[1] · Deyue Yu[1]

Received: 30 March 2020 / Accepted: 30 June 2020

Abstract

As a globally important legume crop, soybean provides excellent sources of protein and oil for human and livestock nutrition. Improving seed protein and oil contents has always been an important objective in soybean breeding. Water-soluble protein plays a significant role in the processing and efficacy of soybean protein. Here, a genome-wide association study (GWAS) of seed compositions (protein, oil, and water-soluble protein contents) was conducted using 211 diverse soybean accessions genotyped with a 355 K SoySNP array. Three, four, and five QTLs were identified related to the protein, oil, and water-soluble protein contents, respectively. Furthermore, five QTLs (*qPC-15-1*, *qOC-8-1*, *qOC-12-1*, *qOC-20-1* and *qWSPC-8-1*) were detected in multiple environments. Analysis of the favorable alleles for oil and water-soluble protein contents showed that *qOC-8-1* (*qWSPC-8-1*) exerted inverse effects on oil and water-soluble protein synthesis. Relative expression analysis suggested that *Glyma.15G049200* in *qPC-15-1* affects protein synthesis and *Glyma.08G107800* in *qOC-8-1* and *qWSPC-8-1* might be involved in oil and water-soluble protein synthesis, producing opposite effects. The candidate genes and significant SNPs detected in the present study will allow a deeper understanding of the genetic basis for the regulation of protein, oil and water-soluble protein contents and provide important information that could be utilized in marker-assisted selection for soybean quality improvement.

Keywords Soybean · Protein · Oil · Water-soluble protein · Genome-wide association

Communicated by Stefan Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-020-01704-7) contains supplementary material, which is available to authorized users.

✉ Guizhen Kan
kanguizhen@njau.edu.cn

✉ Deyue Yu
dyyu@njau.edu.cn

[1] National Key Laboratory of Crop Genetics and Germplasm Enhancement, National Center for Soybean Improvement, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing 210095, China

[2] Jiangsu Yanjiang Institute of Agricultural Sciences, Nantong 226000, China

[3] Collaborative Innovation Center of Henan Grain Crops, College of Agronomy, Henan Agricultural University, Zhengzhou 450000, China

[4] School of Life Sciences, Guangzhou University, Guangzhou 510006, China

Introduction

Soybean (*Glycine max* L. Merr) is an important economic crop because it contains approximately 40% protein content (PC) and 20% oil content (OC), which supplies a great source of human food, cooking oil, and livestock feed, and has biodiesel production, industrial, and pharmaceutical applications (Chaudhary et al. 2015; Clemente and Cahoon 2009). The demand for soy food production has increased fivefold as a result of the awareness of the nutritional value and health benefits of soybean in the human diet (Patil et al. 2017). As an important source of edible oils and biodiesel, the OC in soybean is relatively low compared to most other oilseed crops (Singh and Hymowitz 1999). The water-soluble protein content (WSPC) is a critical trait that affects the soybean food quality and the production of isolated proteins (Zhang et al. 2014). Therefore, increased PC, OC, and WSPC in soybean seeds are desirable and would dramatically contribute to meeting growing soybean protein and oil demands. However, information about the mechanism

Published online: 02 October 2020

controlling the variation of the PC, OC, and WSPC is limited. Furthermore, the negative genetic correlations of the PC with the OC in soybean has made the development of soybean lines with high PC and OC through conventional breeding a challenge (Eskandari et al. 2013; Patil et al. 2017).

In recent decades, extensive efforts have been made to dissect the genetic basis of the soybean PC and OC. More than 100 quantitative trait loci (QTLs) have been reported for the two traits (Phansak et al. 2016). However, only a few reports have focused on mapping QTLs associated with the WSPC in soybean (Lu et al. 2013; Zhang et al. 2014, 2017). Most QTLs were detected in the genetic background of biparental populations, and some of the QTLs were inconsistent. In addition, the resolution of QTL mapping via biparental populations is limited by the number of recombination events, which makes it difficult to apply in breeding programs based on marker-assisted selection (MAS) (Li et al. 2018). Genome-wide association studies (GWAS) based on recombination events that occurred during the evolutionary process provide a complementary tool for detecting the underlying complex traits of QTLs. The resolution and accuracy of QTLs detected in a GWAS depend on the characteristics of the association panel, including its genetic diversity, population structure, linkage disequilibrium (LD), and density of markers (Li et al. 2018). Extensive natural variation may facilitate GWAS by overcoming the allele limitations in the parents of segregating populations. GWAS has been applied in different populations to identify QTLs associated with PC, OC, and WSPC. These reports not only confirmed QTLs discovered by linkage mapping, but also identified novel QTLs. Thus far, several genes related to the soybean PC and OC have been cloned and functionally identified, but no WSPC-related genes have been reported. Further surveys are needed to complement the set of QTLs governing the PC, OC, and WSPC.

Gene-pyramiding molecular breeding combined with conventional breeding techniques has become the main strategy in crop breeding. Crop gene-pyramiding molecular breeding requires more favorable alleles or genes. In addition, the expression of a gene is sensitive to the environment, as is the case for the genes controlling the contents of protein and oil in soybean. Although some QTLs and markers have been obtained, new loci and genes need to be found across different environments. Different significant loci and genes with large effects on traits could be detected in different environments by GWAS. Li et al. (2017) discovered loci on chromosomes A04 and A10 in different environments with great effects on Verticillium wilt resistance in upland cotton. In addition to *qPDH1* on chromosome 16, Hu et al. (2019) determined that *Glyma09g06290* in the flank region of significant SNPs on chromosome nine in different environments might be involved in pod dehiscence. Zhao et al. (2019) reported that 16 genes derived form 31 QTNs (quantitative trait nucleotides), which were detected in different environments, were significantly associated with the seed weight in soybean. Based on GWAS, Wang et al. (2020) revealed two novel loci for photosynthesis-related traits in soybean in different environments. In the previous study, a total of 32 significant loci (16 for PC and 16 for WSPC) across 18 of 20 soybean chromosomes significantly associated with the WSPC and PC across multi-environments was obtained by GWAS (Zhang et al. 2017). In addition, 110 SNPs on three chromosomes (8, 12, and 20) were identified as associated with the oil content across at least two environments (Zhang et al. 2019). Based on the GWAS results and gene annotation, several genes related to corresponding traits were identified (Zhang et al. 2017, 2019).

To identify new genes controlling the genetic architecture components of the PC, OC and WSPC, we conducted a GWAS incorporating 211 diverse accessions with a high-resolution single nucleotide polymorphism (SNP) array (NJAU 355K SoySNP array) based on phenotypic data in different environments. In total, 12 QTLs were identified related to the three traits, among which *qOC-8-1* (*qWSPC-8-1*) had opposing effects on the OC and WSPC. Two candidate genes, *Glyma.15G049200* and *Glyma.08G107800*, might be involved in the synthesis of the PC, OC, and WSPC. These genes and SNPs ultimately may be used in MAS programs for the breeding of high-PC, high-OC and high-WSPC lines.

Materials and methods

Plant materials

A population consisting of a diverse collection of 211 soybean accessions was used in this study. This population originated from 25 provinces of China and was distributed across all three ecological habitats of China. In addition, ten of these soybean accessions were overseas cultivars from the USA, Japan, and Brazil (Wang et al. 2016).

All 211 lines were planted over the course of 3 years (2015, 2016 and 2017; designated E1, E2 and E3, respectively) at the experimental station of the Jiangsu Yanjiang Institute of Agricultural Sciences, Nantong, China, in a completely randomized design with three replicates. The seeds were sown in a row with the following dimensions: 2 m length, 0.5 m width, and 0.10 m interplant spacing. After harvesting, seeds from the replicated samples were pooled based on accession and dried to a constant weight in an oven.

Phenotypic data collection

Determination of the soybean PC, OC, and WSPC was conducted with a near-infrared spectrophotometer (NIR) seed analyzer (DA7200, Perten Instruments, Huddinge, Sweden) as described by Zhang et al. (2014). Calibrations of the instrument were performed by Perten Instruments and the Inspection and Testing Center for Quality of Cereals and Products of the Ministry of Agriculture of China. These calibrations involved more than 700 diverse soybean samples with varied PC and OC, and 146 soybean samples had varied WSPC. Approximately. 60 g dry seeds were placed in a 75 mm-diameter cup that rotated during NIRS scanning. Three scans were conducted on each sample, and the data obtained was the average of the three scans. Afterward, the seeds were mixed, and three technical replicates were used for each material.

Statistical analyses of phenotypic data

Descriptive statistics were obtained and correlation analyses were conducted with SPSS version 20.0 (https://www.ibm.com/products/spss-statistics). An analysis of variance (ANOVA) of the phenotypic data was performed with SAS version 9.2 software (https://www.sas.com). The broad-sense heritability (h^2) of every trait was calculated using the following equation:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma_e^2}{re}},$$

where

σ_g^2

is genotypic variance,

σ_{ge}^2

is genotypic × environmental variance,

σ_e^2

is error variance, e is the number of environments, and r is the number of replicates. The value of h^2 was calculated with SAS version 9.2.

Association mapping

A 355K SoySNP array was developed to genotype the 211 soybean accessions used in this study (Wang et al. 2016). A total of 207,608 SNPs with a minor allele frequency (MAF) greater than 0.05 was employed to perform the GWAS. The PC, OC, and WSPC data used in the association mapping approach were the averages of three replicates. The GWAS was performed in the R package GAPIT 3.0 with a compressed mixed linear model (CMLM) (Lipka et al. 2012). A threshold of $1/n$ (n is the number of markers; $P \leq 4.82 \times 10^{-6}$ or $-\log_{10}P \geq 5.32$) was used in this study.

Haplotype analysis

Haploview 4.2 software was utilized to estimate the squared allele frequency correlation coefficient (r^2) of the alleles (Barrett et al. 2004). In addition, the haplotype block was defined by the "Four Gamete Rule" algorithm.

Candidate gene selection and expression analysis

Based on the average LD decay distance of significant SNPs in cultivated soybeans (130 kb) in this population (Wang et al. 2016), we predicted putative genes with reference to the Glyma.Wm82.a2.v1 genome (https://www.soybase.org/), as in previous studies (Li et al. 2017; Si et al. 2016).

We used public gene annotation data sets from Phytozome (https://phytozome.jgi.doe.gov/) and performed BLASTP analysis against Arabidopsis proteins using the amino acid sequences of putative genes to predict gene functions. To ensure that the selected materials were in the same developmental stage, we selected 12 soybean accessions with different PC, OC, and WSPC for qRT-PCR. Seeds were harvested from soybeans at the half-weight stage of the whole fresh seed and stored at −70 °C for RNA isolation (Liu et al. 2014; Song et al. 2013). The total RNA of the seeds was isolated using a kit (TIANGEN, China). A first-strand cDNA synthesis kit (Vazyme, China) was used to synthesize cDNA. With the soybean tubulin gene (GenBank number: AY907703) as an internal control, qRT-PCR was performed using the ABI 7500 system (Applied Biosystems, USA) under the following conditions: denaturation at 95 °C for 5 min, followed by 40 cycles at 95 °C for 15 s and at 60 °C for 60 s. The fold change in the expression level of each gene was calculated as the $2^{-\Delta\Delta CT}$ value. All experiments were performed with three biological and three technical replicates. The specific primers used for the relative expression analysis of candidate genes in soybean seeds are listed in Supplemental Table 1.

Results

The PC, OC, and WSPC display significant variation in soybean seeds

Significant phenotypic variation in the PC, OC, and WSPC of soybean seeds was observed among 211 accessions in

Table 1 Descriptive statistics, ANOVA results, and broad-sense heritability (h^2) of the PC, OC, and WSPC of soybean across three environments

Trait	Mean	SD	Median	Min	Max	CV (%)	Skew	Kur	G[a]	E[b]	G×E[c]	h^2 (%)[d]
PC-E1	42.75	26.38	42.64	36.15	50.10	6.2	0.35	0.28	***	***	***	64.10
PC-E2	44.27	24.02	44.34	38.24	50.30	5.4	−0.06	−0.49				
PC-E3	43.48	21.77	43.41	37.25	48.98	5.0	0.06	−0.03				
OC-E1	15.89	19.61	15.79	10.13	21.27	12.3	−0.04	0.34	***	***	***	92.26
OC-E2	16.37	18.88	16.41	10.99	21.99	11.5	−0.03	0.40				
OC-E3	17.64	17.82	17.47	11.24	23.00	10.1	0.02	1.03				
WSPC-E1	29.41	47.55	30.33	9.52	42.00	16.17	−1.07	2.65	***	***	***	93.95
WSPC-E2	25.07	48.88	25.70	6.85	35.82	19.50	−1.24	3.22				
WSPC-E3	29.05	47.88	29.92	10.25	40.93	16.48	−1.40	3.10				

***Significant at the 0.001 probability level

[a]Genotype

[b]Environment

[c]Genotype × environment

[d]Broad-sense heritability

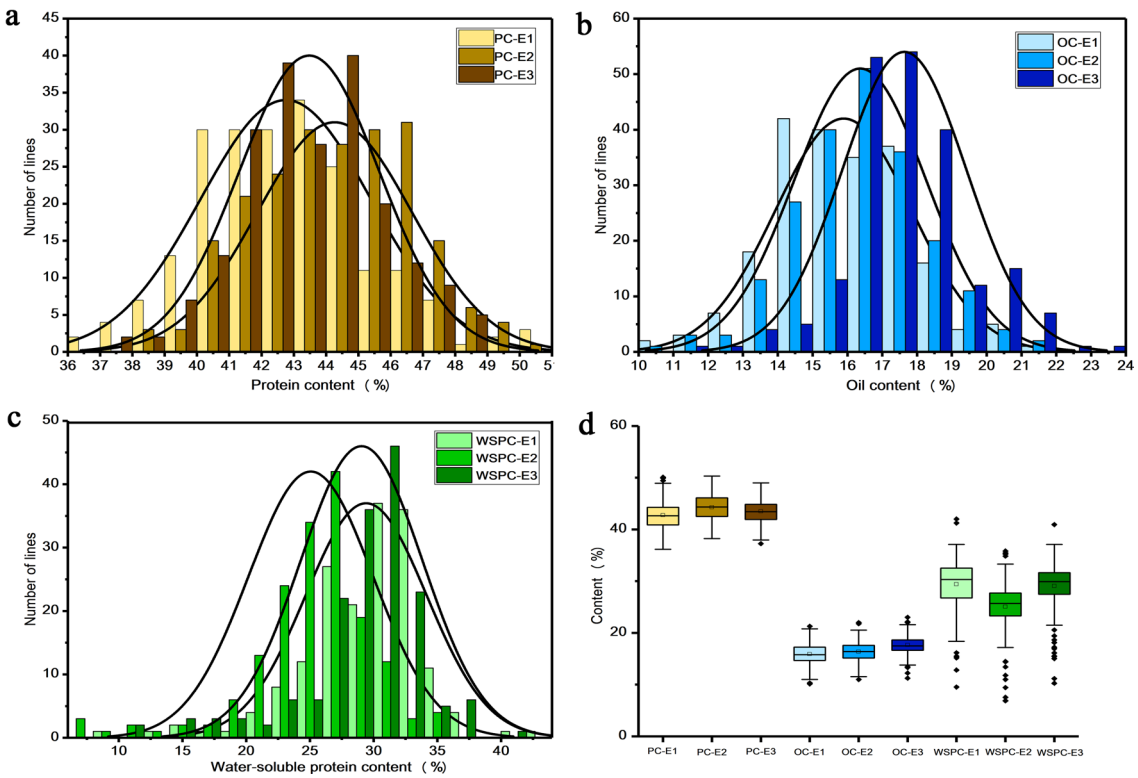

Fig. 1 Frequency distribution of variation for the PC, OC, and WSPC in the 211 association-mapping populations across three environments; **a** the trait of the PC; **b** the trait of the OC; **c** the trait of the WSPC; **d** box plot of the three trait distributions across three environments

this study (Table 1; Fig. 1). For the three traits, the maximum phenotypic values were approximately 1.4-, 2.3-, and 6.1-fold higher than the minimum values across the three environments, respectively (Table 1; Fig. 1). A higher CV (coefficient of variation) was found for the OC and, especially, WSPC compared to that for the PC across the three environments. ANOVA revealed that the genotype and genotype-by-environment interactions significantly influenced the PC, OC, and WSPC ($P < 0.001$), indicating that the three traits were complex traits controlled by multiple factors. In addition, the h^2 values of the PC, OC, and WSPC were 64.10%, 92.26%, and 93.95%, respectively, indicating

that genetic factors play an important role in the formation of the three traits. The correlation analysis of the three traits across the three environments showed that the WSPC had a higher negative relationship with the OC than the PC with the OC (Supplemental Table 2). As a component of the PC, the WSPC had a low correlation coefficient with the PC during the same year, although the correlations were significant at the 0.01 level in all of the environments.

Significant loci controlling the PC, OC, and WSPC

The combined analysis of the three environments was regarded as the fourth environment. A GWAS was performed based on 207,608 SNPs with an MAF > 0.05 across the four environments. According to the significance threshold of $-\log_{10}P \geq 5.32$, in total, 9, 90, and 162 SNPs distributed on chromosomes 6, 7, 8, 11, 12, 13, 15, 17, 19, and 20 were detected for the PC, OC, and WSPC, respectively (Fig. 2; Supplemental Table 3).

AX-93634511 on chromosome 8 was related to both the OC and WSPC in all four environments. Thus, it was a pleiotropic locus for the OC and WSPC. The favorable allele (AX-93634511-G) for the OC was unfavorable for the WSPC, and vice versa (Supplemental Fig. 1). The result was consistent with the significant negative correlation between the two traits. Sixty-four SNPs on chromosome 8 were associated with both the OC and WSPC. Favorable alleles for the OC at these loci were all different from those for the WSPC. The genotypes with favorable alleles at loci related to the OC exhibited high values for the OC and low values for the WSPC, and the genotypes with favorable alleles at loci related to the WSPC exhibited high values for the WSPC and low values for the OC (Supplemental Fig. 2). The results indicated that these 64 SNPs were pleiotropic and had inverse effects in the OC and WSPC.

For one trait, lead SNPs distributed within a region of less than or approximately 2 Mb were considered to be caused by one common gene (Powell et al. 2012). Based on this rule, 12 loci (three for PC, four for OC, and five for WSPC) were identified (Table 2). Of these 12 loci, eight were co-localized with previously reported QTLs for the PC, OC, or related traits. The remaining four (*qOC-12-1*,

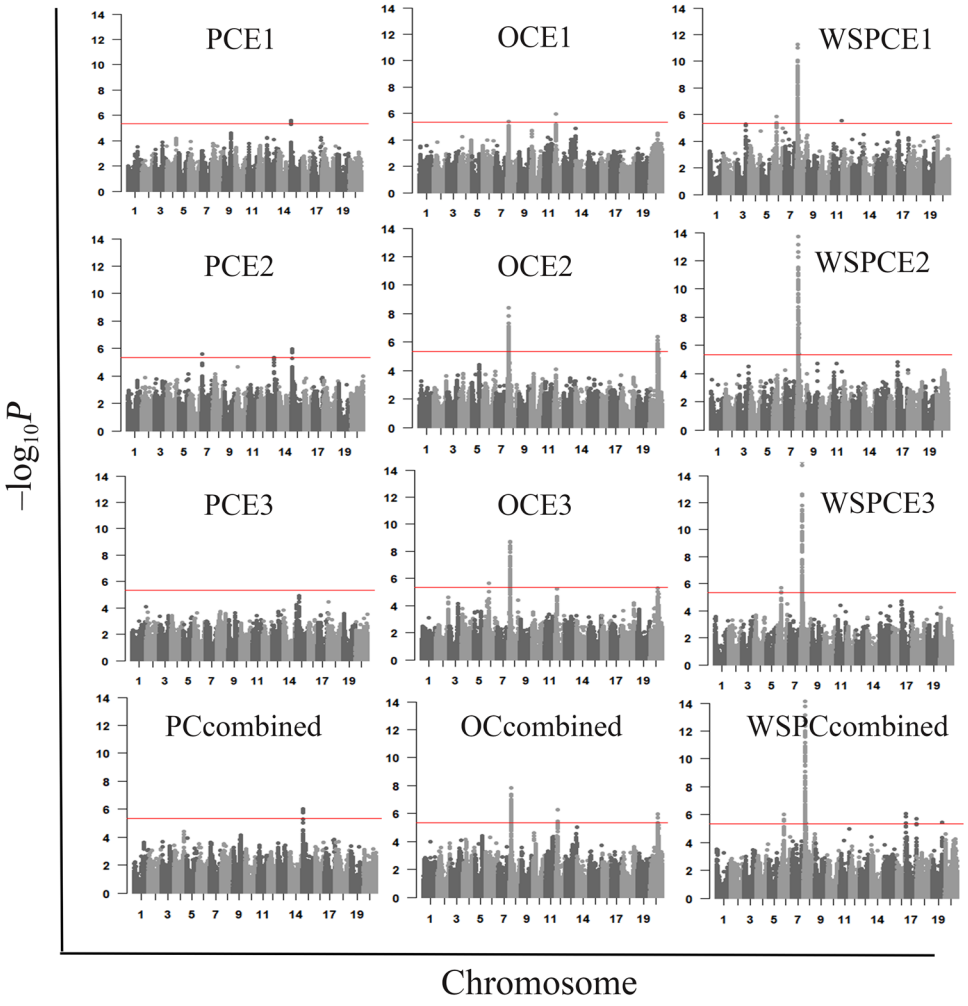

Fig. 2 Manhattan plots of the GWAS results for the PC, OC, and WSPC of soybean seed in three environments and the combined environment. The red lines on the Y-axis designate the significance threshold ($-\log_{10}P = 5.32$). The numbers on the X-axis represent the chromosomes of soybean

Table 2 QTLs associated with PC, OC, and WSPC

QTL	Chr[a]	Rep. SNP[b]	Pos.(bp)[c]	No. sig.[d]	R^2 (%)	Environments	Related QTL[e]	References
qPC-7-1	7	AX-94035476	4210426	1	18.51	E2	*cqSeed protein-006*, *Seed protein 12-4*, *Seed Leu 1-6 and Seed Ser 1-6*	Panthee et al. (2006), Pathan et al. (2013) and Specht et al. (2001)
qPC-13-1	13	AX-93812948	25811574	2	17.92	E2	*Seed protein 36-21* and *Seed protein 36-23*	Mao et al.(2013)
qPC-15-1	15	AX-93834405	3846799	6	33.94	E1, E2, Combined	*Seed protein 30-3*, *Seed protein 39-2 cqSeed protein-008*	Pathan et al. (2013), Phansak et al. (2016), Tajuddin et al. (2003), Warrington et al. (2015) and Zhou et al. (2015)
qOC-6-1	6	AX-93730252	19251651	1	23.51	E3	*Seed oil 27-6*, *Seed oleic 3-1*,*Seed oil 23-1*	Hyten et al. (2004) and Reinprecht et al. (2006)
qOC-8-1	8	AX-94284412	8434697	69	30.13	E1, E2, E3, Combined	*Seed oil 30-2*, *Seed oil 1-1*, *GqOil-8*	Liang et al. (2010), Mansur et al. (1993) and Zhang et al. (2019)
qOC-12-1	12	AX-93799425	9862679	3	26.71	E1, Combined	–	–
qOC-20-1	20	AX-93905778	31641964	17	26.31	E2, Combined	*Seed oil 2-1*, *Seed oleic 6-11*,*Seed oil 15-1*	Bachlava et al. (2009), Chung et al. (2003) and Diers et al. (1992)
qWSPC-8-1	8	AX-93930669	8276381	141	56.24	E1, E2, E3, Combined	*Seed protein 34-4*, *Seed protein 26-1*, *cqSeed protein-016*, *GqWSPC-8*	Lu et al. (2013), Pathan et al. (2013), Reinprecht et al. (2006) and Zhang et al. (2017)
qWSPC-11-1	11	AX-94091850	32951114	1	23.12	E1	*Seed protein 26-6*	Reinprecht et al. (2006)
qWSPC-17-1	17	AX-93856909	959379	17	26.06	Combined	–	–
qWSPC-17-2	17	AX-93867540	40976006	2	25.14	Combined	–	–
qWSPC-19-1	19	AX-93658470	37860368	1	24.35	Combined	–	–

[a]Chromosome

[b]The representative SNP with the min *P* value

[c]Physical position

[d]The number of significant SNPs detected in the region

[e]Reported QTLs related to the PC, OC, and WSPC or any other related traits collected from SoyBase

qWSPC-17-1, *qWSPC-17-2*, and *qWSPC-19-1)* were novel QTLs. Among the reported QTLs, *qOC-8-1* for the OC was also associated with the WSPC in multiple environments. This QTL has been detected to be associated with the PC and OC by linkage mapping and association mapping (Liang et al. 2010; Lu et al. 2013; Mansur et al. 1993; Pathan et al. 2013; Reinprecht et al. 2006; Zhang et al. 2017, 2019). *qPC-15-1*, which is a unique QTL located on chromosome 15 that was associated with the PC in multiple environments, was previously detected to be associated with the PC in a recombinant inbred line (RIL) population and six BC_1F_9-derived families (Kim et al. 2016; Tajuddin et al. 2003).

Candidate genes associated with the PC, OC, and WSPC

To identify candidate genes with strong effects on traits and reduce the false-positive rate, we investigated candidate genes in the 130-kb flanking regions upstream and downstream of the SNPs repeatedly identified in multiple environments or with low *P* values. AX-93834405 related to the PC in *qPC-15-1* and AX-93634511 related to the OC in *qOC-8-1* were detected repeatedly in most environments. AX-93930669 with minimum *P* values was related to the WSPC in the combined analysis of the three environments.

Glyma.15G049200 and *Glyma.15G049800* were located in the flanking region of AX-93834405 in *qPC-15-1*. *Glyma.15G049200* is a member of the SWEET (Sugars Will Eventually be Exported Transporters) family, which is responsible for transporting sucrose molecules across a membrane (Chen et al. 2012; Patil et al. 2015). Smith et al. (1989) reported that sucrose in soybean seeds could be metabolized to produce precursors for protein. Seed development was accompanied by the accumulation of storage substances. Due to *Glyma.15G049200* being involved in the transportation of carbohydrates and its high expression during the process of soybean seed development (Supplemental Table 6), it was selected as a candidate gene for the PC in the expression analysis. *Glyma.15G04980* belongs to the nodulin MtN21 transporter protein family, which is essential for effective symbiotic N_2 fixation, nodule development, and NH_4^+ transport. Nitrogen resources are important for the accumulation of protein. *Glyma.08G107800* was located in the 130-kb flanking regions upstream and downstream of AX-93634511 and AX-93930669. Interestingly, AX-93751849, which was located in the same block as AX-93930669, causes an amino acid change in *Glyma.08G107800* (Supplemental Table 4; Supplemental Fig. 3). *Glyma.08G107800* encodes a bifunctional aspartate kinase/homoserine dehydrogenase (AK-HSDH), which is important for the synthesis of amino acids in the aspartic acid family (Lys, Thr, Met, and Ile). Amino acids in the aspartic acid family are the major agents involved in the biosynthesis of other amino acids. Furthermore, in *Arabidopsis*, sucrose-stimulated *AK-HSDH* expression to facilitate the utilization of sucrose for Asp-related amino acids biosynthesis and the synthesis of fatty acids was sensitive to the availability of sucrose (Lam et al., 1994; Zhu-Shimoni and Galili 1998). Storage compound synthesis (mainly protein and oil in soybean) in a seed includes the processes of carbon source competition and allocation. We speculated *Glyma.08G107800* might regulate the synthesis of oil by affecting the amount of available sucrose. Thus, we selected it as a candidate gene related to the OC and WSPC.

Expression analysis of candidate genes related to the PC, OC, and WSPC

To elucidate the potential functions of the three genes, qRT-PCR analysis was performed to investigate their expression patterns in different accessions with corresponding phenotypes.

We detected the expression of *Glyma.15G049200* and *Glyma.15G049800* in the seeds of 12 soybean accessions (Fig. 3). The 12 accessions represented varieties with high and low PCs in soybean seeds (Fig. 3a). The results of the expression analysis revealed that the relative expression of *Glyma.15G049200*, rather than that of *Glyma.15G049800*, in most high-PC accessions was higher than that in low-PC accessions (Fig. 3b, c). The relative expression result for *Glyma.15G049200* in the selected accessions was generally consistent with the phenotypic variation in the PC in these accessions, which suggested a potential role of this gene in protein synthesis or accumulation in soybean seeds.

We also selected 12 accessions with different OC and WSPC values (Fig. 4a, b). The relative expression of *Glyma.08G107800* in high-OC accessions was lower than that in low-OC accessions (Fig. 4c). Based on the results in which the correlation coefficients between OC and WSPC were significant and negative and *qOC-8-1* (*qWSPC-8-1*) exhibited inverse functions in oil and water-soluble protein synthesis, the expression of *Glyma.08G107800* in high-WSPC accessions was correspondingly higher than that in low-WSPC accessions. Therefore, *Glyma.08G107800* might be negatively related to the synthesis of oils and positively related to water-soluble protein synthesis in soybean.

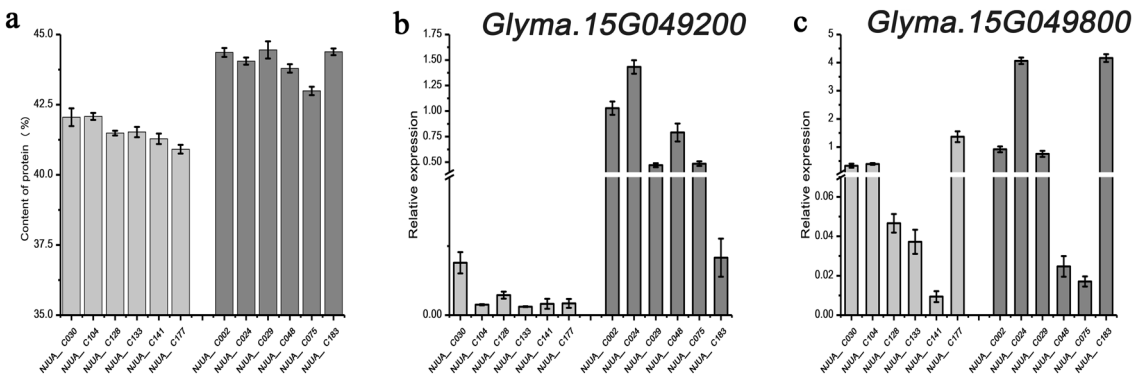

Fig. 3 Protein contents of 12 accessions (**a**) and expression patterns of *Glyma.15G049200* (**b**) and *Glyma.15G049800* (**c**) in the seeds of 12 accessions. The bars represent the standard errors of three technical replicates for three biological replicates

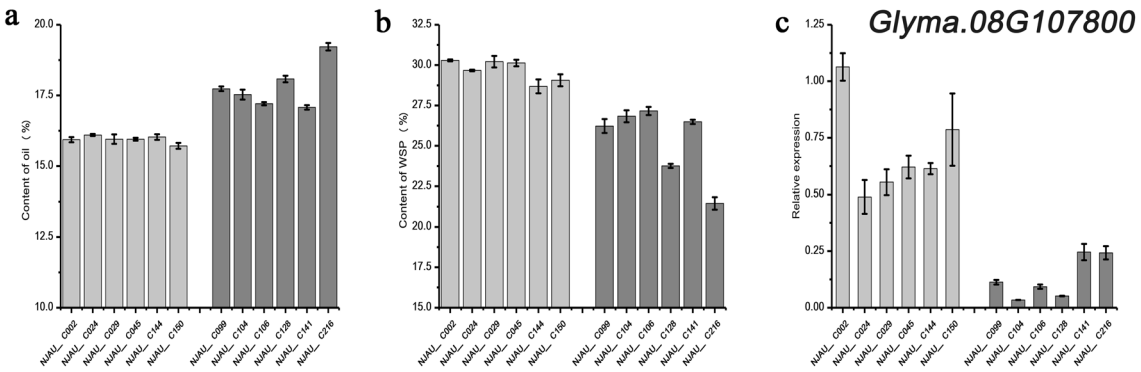

Fig. 4 Content of oil (**a**) and water-soluble protein (**b**) in 12 accessions and expression patterns of *Glyma.08G107800* (**c**) in the seeds of 12 accessions. Bars represent the standard errors of three technical replicates for three biological replicates

Discussion

Soybean quality improvement has always been an important target in breeding programs. As the basic compositions of quality traits in soybean, the PC, OC, and WSPC are complex quantitative traits, which are controlled by multiple genes with minor effects. Furthermore, knowledge about the regulating mechanism of the variation of the PC, OC, and WSPC was previously limited. Based on a GWAS with 355K SoySNP array and 211 collections, a total of 12 QTLs was identified as being related to the three traits with five detected in multiple environments. Two candidate genes might be involved in the pathways of protein, oil, and water-soluble protein synthesis. The significant SNPs and candidate genes detected in this study will provide enlightenment and useful information for the improvement of soybean quality traits.

Phenotypic variation and the relationships of the PC, OC, and WSPC

Given the important roles that the protein and oil in soybean seed play in our daily life, it is of great practical significance to increase the PC, OC, and WSPC in soybean. In this study, the three traits exhibited significant phenotypic variation among 211 accessions, especially for the WSPC. The range of variation of the WSPC was similar to that reported by Zhang et al. (2017) and larger than that reported by Lu et al. (2013). The significant phenotypic variation in the accessions used for association mapping promotes the identification of more relevant markers in a broad genetic background. The heritability of the PC and OC detected in the present study was consistent with the estimated range (PC from 0.57 to 0.99 and OC from 0.51 to 0.99) summarized by Kim et al. (2016). In addition, the heritability of the WSPC was also high. The heritability of the three traits in this study indicated that genetics was the main factor underlying the phenotypic variation. It was possible to detect genetic factors that controlled phenotypic variation in this study. The differences in descriptive statistics for the three traits between the current study and previous studies reconfirmed the quantitative genetic nature and environmental sensitivities of the three traits.

The relationship between the PC and OC found in this study confirmed the negative correlation reported in the previous studies (Li et al. 2018; Patil et al. 2017). The WSPC had a low correlation with the PC in the current study. Thus, the genetic loci that control the PC and WSPC might be different. Unfortunately, the relationship between the OC and WSPC was significant and negative, and the correlation coefficient was high. The strong negative correlation between the OC and WSPC may be one of the reasons for the negative correlation between the OC and PC. In addition, this result suggests the possibility of tightly linked QTLs with opposite pleiotropic effects on the two traits, which will make it difficult for breeders to develop soybean cultivars with a high OC and a high WSPC.

Confirmation of stable QTLs and identification of novel QTLs associated with the PC, OC, and WSPC in soybean

The PC, OC, and WSPC are complex, quantitative, heritable traits controlled by multiple genes (Li et al. 2019; Lu et al. 2013). Many efforts have been made to discover the genetic mechanism underlying these traits. Previous studies used different mapping methods and different populations to detect QTLs associated with the PC and OC, which were distributed on the 20 chromosomes of soybean (https://www.soybase.org/). In the present study, all three QTLs related to the PC were consistent with those in previous reports, which indicated the stability of the QTLs. *qPC-15-1* was designated as an officially confirmed QTL, which overlapped with the results of many different studies involving mapping

with different populations and methods (Pathan et al. 2013; Phansak et al. 2016; Tajuddin et al. 2003; Warrington et al. 2015; Zhou et al. 2015). Kim et al. (2016) mapped this QTL in a 535 kb interval between BARCSOYSSR_15_0161 and BARCSOYSSR_15_0194 with six BC_1F_9-derived families. Van and McHale (2017) performed a meta-QTL analysis to a large number of QTLs related to the protein and oil contents and composition-related QTLs. In total, five meta-QTLs were detected on chromosome 15, and *qPC-15-1* was consistent with *mPO15-2* (117.8 kb) between BARC-008231-00112 and BARC-042857-08439. This QTL was stable across different mapping populations and environments and thus should be used in molecular breeding to improve the soybean protein quality. Compared with the results of previous study, all three QTLs related to the PC were novel QTLs (Zhang et al. 2017).

QTLs linked to the OC in this study were mainly distributed on chromosomes 6, 8, 12, and 20. *qOC-6-1*, *qOC-8-1* and *qOC-20-1* overlapped with QTLs related to oil and its compositional content in different populations, as determined by different mapping methods (Bachlava et al. 2009; Chung et al. 2003; Diers et al. 1992; Hyten et al. 2004; Liang et al. 2010; Mansur et al. 1993; Reinprecht et al. 2006). *qOC-12-1* was a novel QTL identified in E1 and the combined environments that had never been reported in other studies. *qOC-8-1* overlapped with *GqOil-8*, which was detected in a previous study (Zhang et al. 2019). However, it was not the QTL with the greatest effect on the OC in the previous study, unlike in the present study, and the other three QTLs (*qOC-6-1*, *qOC-12-1*, and *qOC-20-1*) were newly discovered in the present study. Two QTLs (*qWSPC-8-1* and *qWSPC-11-1*) associated with the WSPC overlapped with previously reported QTLs related to the PC or WSPC (Lu et al. 2013; Pathan et al. 2013; Reinprecht et al. 2006). In addition, *qWSPC-8-1* overlapped with *GqWSPC-8* identified in a previous study (Zhang et al. 2017). The remaining three QTLs (*qWSPC-17-1, qWSPC-17-2* and *qWSPC-19-1*) were novel QTLs related to the WSPC. Therefore, a GWAS carried out in different environments yielded new results and confirmed our previous results, which suggested the necessity and significance of increasing environments.

In the present study, we elaborated the inverse effects of *qOC-8-1* (*qWSPC-8-1*) on the OC and WSPC in terms of the phenotype, favoring alleles, and expression levels of candidate genes. Although this locus was stable and prevalent, it cannot be utilized in the development of elite cultivars with both high OC and high WSPC. Nevertheless, it can provide useful information for subsequent molecular biological research. This region has also been associated with other traits in soybean, such as resistance to soybean cyst nematodes, salt tolerance, seed coat color, seed weight, internode length, and some quality traits (Guo et al. 2006; Lu et al. 2013; Mansur et al. 1993; Oyoo et al. 2011; Wu et al. 2009). Thus, this locus might be an important genetic region preserved over evolution.

Potential candidate genes for the PC, OC, and WSPC

To date only a few genes related to the PC and OC in soybean have been reported, such as *Gy1* to *Gy7* (encoding glycinin subunits), *CG1* to *CG15* (encoding β-conglycinin subunits) (Li and Zhang 2011), *GsOAS-TL1* (Zhang et al. 2008), *GmOASTL4* (Ning et al. 2010) related to the PC, and *GmbZiP123* (Song et al. 2013), *GmMYB73* (Liu et al. 2014), *GmDof1*/*GmDof4* (Wang et al. 2007) and *GmOLEO1*(Zhang et al. 2019) related to the OC. In the current study, we selected three candidate genes based on the GWAS results. In addition to these three genes, we also detected that genes in *qPC15-1* and *qOC8-1*(*qWSPC-8-1*) *might regulate the PC, OC, and WSPC*, including *Glyma.15g049100*, *Glyma.15g050100* in *qPC-15-1* and *Glyma.08G109200* to *Glyma.08G109500*, *Glyma.08G110300* to *Glyma.08G110500* in *qOC-8-1*(*qWSPC-8-1*), which are involved in the metabolic pathways of proteins, oils and related substances (Supplemental Table 5) (Van and McHale 2017; Xuan et al. 2018). Van and McHale (2017) reported that *Glyma.15g049100* and *Glyma.15g050100* might be involved in carbon partitioning and further regulating the PC and OC in soybean. *Glyma.15G049700* and *Glyma.15G049900* in *qPC-15-1* are also members of the nodulin MtN21 transporter protein family. However, the expression levels of *Glyma.15G049700* and *Glyma.15G049900* in the different kinds of soybean tissues were too low to be detected (Supplemental Table 6). We provided some evidence in terms of the expression levels of genes, and the specific genes regulating these traits need to be verified according to well-designed experiments.

Some genes located in the reported and novel QTLs might also control the PC, OC and WSPC (Supplemental Table 5). For example, *Glyma.07G048300* in *qPC-7-1* encodes a pyrophosphorylase, which plays an important role in the process of glycolysis. Seed-preferred RNA interference-mediated silencing of *PPa1*, which is the homolog of *Glyma.07G048300* in *Arabidopsis*, led to an oil increase of 1–4%, mostly at the expense of seed storage proteins (Meyer et al. 2012). Glycolysis provides precursors for protein and oil synthesis; thus, there is a high probability that genes encoding pyrophosphorylase also regulated the PC and the contents of related materials. *Glyma.13G144900* in *qPC-13-1* encodes an isocitrate dehydrogenase, which is one of the major enzymes in TAC (tricarboxylic acid cycle). Fait et al. (2011) reported that TCA was closely related to the accumulation of storage materials in seeds. *Glyma.11G234600* in *qWSPC-11-1* encodes an amino acid transporter, which might regulate the PC (Cheng et al. 2016). Four genes (*Glyma.17G011500* encoding an aspartyl

protease and *Glyma.19G120400* to *Glyma.19G120600* encoding isopropylmalate synthases) were located in two novel QTLs (*qWSPC-17-1* and *qWSPC-19-1*) also involved in the progress of protein synthesis. A functional study of *AT1G62290*, which is the homolog of *Glyma.17G011500* in *qWSPC-17-1*, showed that it plays an important role in the processing of storage proteins into chains competent for stable accumulation in mature seeds (Gruis et al. 2002). The three genes (*Glyma.19G120400*, *Glyma.19G120500*, and *Glyma.19G120600*) in *qWSPC-19-1* are involved in the synthesis of Leu and related materials (de Kraker and Gershenzon 2011). GWAS is an effective tool with which to detect genes that control quantitative traits like the PC, OC, and WSPC. The QTLs and candidate genes detected in the present study could be useful for the future MAS and the elaboration of the underlying mechanisms of controlling the PC, OC, and WSPC.

Acknowledgements This work was supported in part by the Ministry of Science and Technology (2017YFE0111000), the Fundamental Research Funds for the Central Universities (KYZ201705), the National Natural Science Foundation of China (31571688, 31871649, 31671715), and the Ministry of Agriculture (2016ZX08004-003, 2016ZX08009003-004).

Author contributions GK, DY, and SZ designed the research, and SZ, DH, SZ, and DZ performed the research. SZ, HW, and HD analyzed the data, and SZ wrote the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies involving human participants or animals performed by any of the authors.

References

Bachlava E, Dewey RE, Burton JW, Cardinal AJ (2009) Mapping and comparison of quantitative trait loci for oleic acid seed content in two segregating soybean populations. Crop Sci 49:433–442

Barrett JC, Fry B, Maller J, Daly MJ (2004) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

Chaudhary J, Patil G, Sonah H, Deshmukh R, Vuong T, Valliyodan B, Nguyen H (2015) Expanding omics resources for improvement of soybean seed composition traits. Front Plant Sci 6:1021

Chen LQ, Qu XQ, Hou BH, Sosso D, Osorio S, Fernie AR, Frommer WB (2012) Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. Science 335:207–211

Cheng L, Yuan HY, Ren R, Zhao SQ, Han YP, Zhou QY, Ke DX, Wang YX, Wang L (2016) Genome-wide identification, classification, and expression analysis of amino acid transporter gene family in *Glycine Max*. Front Plant Sci 7:515

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053–1067

Clemente TE, Cahoon EB (2009) Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol 151:1030–1040

de Kraker JW, Gershenzon J (2011) From amino acid to glucosinolate biosynthesis: protein sequence changes in the evolution of methylthioalkylmalate synthase in *Arabidopsis*. Plant Cell 23:38–53

Diers BW, Keim P, Fehr W, Shoemaker R (1992) RFLP analysis of soybean seed protein and oil content. Theor Appl Genet 83:608–612

Eskandari M, Cober ER, Rajcan I (2013) Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theor Appl Genet 126:1677–1687

Fait A, Nesi AN, Angelovici R, Lehmann M, Pham PA, Song L, Haslam RP, Napier JA, Galili G, Fernie AR (2011) Targeted enhancement of glutamate-to-γ-aminobutyrate conversion in arabidopsis seeds affects carbon-nitrogen balance and storage reserves in a development-dependent manner. Plant Physiol 157:1026–1042

Gruis D, Selinger DA, Curran JM, Jung R (2002) Redundant proteolytic mechanisms process seed storage proteins in the bbsence of seed-type members of the vacuolar processing enzyme family of cysteine proteases. Plant Cell 14:2863–2882

Guo B, Sleper DA, Nguyen HT, Arelli PR, Shannon JG (2006) Quantitative trait loci underlying resistance to three soybean cyst nematode populations in soybean PI 404198A. Crop Sci 46:224–233

Hu D, Kan G, Hu W, Li Y, Hao D, Li X, Yang H, Yang Z, He X, Huang F, Yu D (2019) Identification of loci and candidate genes responsible for pod dehiscence in soybean via genome-wide association analysis across multiple environments. Front Plant Sci 10:811

Hyten DL, Pantalone VR, Sams CE, Saxton AM, Landau-Ellis D, Stefaniak TR, Schmidt ME (2004) Seed quality QTL in a prominent soybean population. Theor Appl Genet 109:552–561

Kim M, Schultz S, Nelson RL, Diers BW (2016) Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci 56:219–225

Lam HM, Peng SS, Coruzzi GM (1994) Metabolic regulation of the gene encoding glutamine-dependent asparagine synthetase in *Arabidopsis thaliana*. Plant Physiol 106:1347–1357

Li C, Zhang YM (2011) Molecular evolution of glycinin and beta-conglycinin gene families in soybean (*Glycine max* L. Merr.). Heredity 106:633–641

Li T, Ma X, Li N, Zhou L, Liu Z, Han H, Gui Y, Bao Y, Chen J, Dai X (2017) Genome-wide association study discovered candidate genes of Verticillium wilt resistance in upland cotton (*Gossypium hirsutum L.*). Plant Biotechnol J 15:1520–1532

Li YH, Reif JC, Hong HL, Li HH, Liu ZX, Ma YS, Li J, Tian Y, Li YF, Li WB, Qiu LJ (2018) Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. Plant Sci 266:95–101

Li S, Xu H, Yang J, Zhao T (2019) Dissecting the genetic architecture of seed protein and oil content in soybean from the Yangtze and Huaihe River Valleys using multi-locus genome-wide association studies. Int J Mol Sci 20:3041

Liang HZ, Yu YL, Wang SF, Lian Y, Wang TF, Wei YL, Gong PT, Liu XY, Fang XJ, Zhang MC (2010) QTL mapping of isoflavone, oil and protein contents in soybean (*Glycine max* L. Merr.). Ag Sci China 9:1108–1116

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399

Liu YF, Li QT, Lu X, Song QX, Lam SM, Zhang WK, Ma B, Lin Q, Man WQ, Du WG, Shui GH, Chen SY, Zhang JS (2014) Soybean *GmMYB73* promotes lipid accumulation in transgenic plants. BMC Plant Biol 14:73

Lu W, Wen Z, Li H, Yuan D, Li J, Zhang H, Huang Z, Cui S, Du W (2013) Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. Theor Appl Genet 126:425–433

Mansur L, Lark K, Kross H, Oliveira A (1993) Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). Theor Appl Genet 86:907–913

Mao T, Jiang Z, Han Y, Teng W, Zhao X, Li W (2013) Identification of quantitative trait loci underlying seed protein and oil contents of soybean across multi-genetic backgrounds and environments. Plant Breed 132:630–641

Meyer K, Stecca KL, Ewell-Hicks K, Allen SM, Everard JD (2012) Oil and protein accumulation in developing seeds is influenced by the expression of a cytosolic pyrophosphatase in *Arabidopsis*. Plant Physiol 159:1221–1234

Ning H, Zhang C, Yao Y, Yu D (2010) Overexpression of a soybean *O*-acetylserine (thiol) lyase-encoding gene *GmOASTL4* in tobacco increases cysteine levels and enhances tolerance to cadmium stress. Biotechnol Lett 32:557–564

Oyoo ME, Benitez ER, Kurosaki H, Ohnishi S, Miyoshi T, Kiribuchi-Otobe C, Horigane A, Takahashi R (2011) QTL analysis of soybean seed coat discoloration associated with *II TT* genotype. Crop Sci 51:464–469

Panthee DR, Pantalone VR, Saxton AM, West DR, Sams CE (2006) Genomic regions associated with amino acid composition in soybean. Mol Breed 17:79–89

Pathan SM, Vuong T, Clark K, Lee J-D, Shannon JG, Roberts CA, Ellersieck MR, Burton JW, Cregan PB, Hyten DL, Nguyen HT, Sleper DA (2013) Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. Crop Sci 53:765–774

Patil G, Valliyodan B, Deshmukh R, Prince S, Nicander B, Zhao M, Sonah H, Song L, Lin L, Chaudhary J, Liu Y, Joshi T, Xu D, Nguyen HT (2015) Soybean (*Glycine max*) SWEET gene family: insights through comparative genomics, transcriptome profiling and whole genome re-sequence analysis. BMC Genom 16:520

Patil G, Mian R, Vuong T, Pantalone V, Song Q, Chen P, Shannon GJ, Carter TC, Nguyen HT (2017) Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor Appl Genet 130:1975–1991

Phansak P, Soonsuwon W, Hyten D, Song Q, Cregan P, Graef G, Specht J (2016) Multi-population selective genotyping to identify soybean [*Glycine max* (L.) Merr.] seed protein and oil QTLs. G3 (Bethesda) 6:1635–1648

Powell JE, Henders AK, McRae AF, Caracella A, Smith S, Wright MJ, Whitfield JB, Dermitzakis ET, Martin NG, Visscher PM, Montgomery GW (2012) The brisbane systems genetics study: genetical genomics meets complex trait genetics. PLoS ONE 7:e35430

Reinprecht Y, Poysa VW, Yu K, Rajcan I, Ablett GR, Pauls KP (2006) Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. Genome 49:1510–1527

Si L, Chen J, Huang X, Gong H, Luo J, Hou Q, Zhou T, Lu T, Zhu J, Shangguan Y, Chen E, Gong C, Zhao Q, Jing Y, Zhao Y, Li Y, Cui L, Fan D, Lu Y, Weng Q, Wang Y, Zhan Q, Liu K, Wei X, An K, An G, Han B (2016) *OsSPL13* controls grain size in cultivated rice. Nat Genet 48:447–456

Singh RJ, Hymowitz T (1999) Soybean genetic resources and crop improvement. Genome 42:605–616

Smith AJ, Rinne RW, Seif RD (1989) Phosphoenolpyruvate carboxylase and pyruvate kinase involvement in protein and oil biosynthesis during soybean seed development. Crop Sci 29:349–353

Song QX, Li QT, Liu YF, Zhang FX, Ma B, Zhang WK, Man WQ, Du WG, Wang GD, Chen SY, Zhang JS (2013) Soybean *GmbZIP123* gene enhances lipid content in the seeds of transgenic *Arabidopsis* plants. J Exp Bot 64:4329–4341

Specht J, Chase K, Macrander M, Graef G, Chung J, Markwell J, Germann M, Orf J, Lark K (2001) Soybean response to water: a QTL analysis of drought tolerance. Crop Sci 41:493–509

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. Breed Sci 53:133–140

Van K, McHale LK (2017) Meta-analyses of QTLs associated with protein and oil contents and compositions in soybean [*Glycine max* (L.) Merr.] seed. Int J Mol Sci 18:1180

Wang HW, Zhang B, Hao YJ, Huang J, Tian AG, Liao Y, Zhang JS, Chen SY (2007) The soybean Dof-type transcription factor genes, *GmDof4* and *GmDof11*, enhance lipid content in the seeds of transgenic Arabidopsis plants. Plant J 52:716–729

Wang J, Chu S, Zhang H, Zhu Y, Cheng H, Yu D (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. Sci Rep 6:20728

Wang L, Yang Y, Zhang S, Che Z, Yuan W, Yu D (2020) GWAS reveals two novel loci for photosynthesis-related traits in soybean. Mol Genet Genom. https://doi.org/10.1007/s00438-020-01661-1

Warrington CV, Abdel-Haleem H, Hyten DL, Cregan PB, Orf JH, Killam AS, Bajjalieh N, Li Z, Boerma HR (2015) QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor Appl Genet 128:839–850

Wu X, Blake S, Sleper D, Shannon J, Cregan P, Nguyen H (2009) QTL, additive and epistatic effects for SCN resistance in PI 437654. Theor Appl Genet 118:1093–1105

Xuan L, Zhang C, Yan T, Wu D, Hussain N, Li Z, Chen M, Pan J, Jiang L (2018) *TRANSPARENT TESTA 4*-mediated flavonoids negatively affect embryonic fatty acid biosynthesis in *Arabidopsis*. Plant Cell Environ 41:2773–2790

Zhang C, Meng Q, Gai J, Yu D (2008) Cloning and functional characterization of an *O*-acetylserine(thiol)lyase-encoding gene in wild soybean (*Glycine soja*). Mol Biol Rep 35:527–534

Zhang D, Kan G, Hu Z, Cheng H, Zhang Y, Wang Q, Wang H, Yang Y, Li H, Hao D, Yu D (2014) Use of single nucleotide polymorphisms and haplotypes to identify genomic regions associated with protein content and water-soluble protein content in soybean. Theor Appl Genet 127:1905–1915

Zhang D, Lu H, Chu S, Zhang H, Zhang H, Yang Y, Li H, Yu D (2017) The genetic architecture of water-soluble protein content and its genetic relationship to total protein content in soybean. Sci Rep 7:5053

Zhang D, Zhang H, Hu Z, Chu S, Yu K, Lv L, Yang Y, Zhang X, Chen X, Kan G, Tang Y, An YC, Yu D (2019) Artificial selection on *GmOLEO1* contributes to the increase in seed oil during soybean domestication. PLoS Genet 15:e1008267

Zhao X, Dong H, Chang H, Zhao J, Teng W, Qiu L, Li W, Han Y (2019) Genome wide association mapping and candidate gene analysis for hundred seed weight in soybean [*Glycine max* (L.) Merrill]. BMC Genom 20:648

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T, Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B,

Lee S-H, Wang W, Tian Z (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:408–414

Zhu-Shimoni JX, Galili G (1998) Expression of an arabidopsis aspartate kinase/homoserine dehydrogenase gene is metabolically regulated by photosynthesis-related signals but not by nitrogenous compounds. Plant Physiol 116:1023–1028