Theor Appl Genet (2014) 127:1905–1915
DOI 10.1007/s00122-014-2348-1

ORIGINAL PAPER

Use of single nucleotide polymorphisms and haplotypes to identify genomic regions associated with protein content and water-soluble protein content in soybean

Dan Zhang · Guizhen Kan · Zhenbin Hu · Hao Cheng · Yu Zhang · Qing Wang · Hui Wang · Yuming Yang · Hongyan Li · Derong Hao · Deyue Yu

Received: 20 March 2014 / Accepted: 7 June 2014 / Published online: 21 June 2014

Abstract

Key message **Four major SPC-specific loci were identified, and these accounted for 8.5–15.1 % of the phenotypic variation, thus explaining why certain soybean varieties have a high PC but a low SPC.**

Abstract Water-soluble protein content (SPC) is a critical factor in both food quality and the production of isolated soybean proteins. However, few data are available regarding the genetic control and the mechanisms contributing to elevated SPC. In this study, a soybean collection of 192 accessions from a wide geographic range was used to identify genomic regions associated with soybean protein content (PC) and SPC using an association mapping approach employing 1,536 SNP makers and 232 haplotypes. The diverse panel revealed a large genetic variation in PC and SPC. Association mapping was performed using three methods to minimize false-positive associations. This resulted in 4/8 SNPs and 3/6 haplotypes that were significantly associated with soybean PC/SPC in two or more environments based on the mixed model. An SNP that was highly significantly associated with PC, BARC-021267-04016, was localized 0.28 cM away from a published glycinin gene, *G7*, and was detected across all four environments. Four major SPC-specific loci, BARC-029149-06088, BARC-018023-02499, BARC-041663-08059 and haplotype 15 (hp15), were stably identified on chromosomes five and eight and explained 8.5–15.1 % of the phenotypic variation. Moreover, a glutelin type-B 2-like gene was identified on chromosome eight and may be related to soybean protein solubility. These markers, which are located in previously reported QTL, reconfirmed previous findings and may be important targets for the identification of protein-related genes. These novel SNPs and haplotypes are important for further understanding the genetic basis of PC and SPC. In addition, by comparing the correlation and genetic loci between PC and SPC, we provide new insights into why certain soybean varieties have a high protein content but a low SPC.

Introduction

Soybean is the only vegetable food that contains complete protein, as it provides all the essential amino acids for human nutrition (Erdman 2000). Soybean proteins are used as functional and nutritional ingredients in a wide variety of food products and as a substitute for animal-derived proteins (Singh et al. 2008). In actual production, processed soybean protein appears in foods mainly in three forms:

Communicated by Istvan Rajcan.

D. Zhang and G. Kan contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2348-1) contains supplementary material, which is available to authorized users.

D. Zhang (✉) · Y. Zhang · Y. Yang · H. Li
Collaborative Innovation Center of Henan Grain Crops, College of Agronomy, Henan Agricultural University, Zhengzhou 450002, China
e-mail: zhangdan8006@163.com

G. Kan · Z. Hu · H. Cheng · Q. Wang · H. Wang · D. Yu
National Center for Soybean Improvement, National Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China

D. Yu
e-mail: dyyu@njau.edu.cn

D. Hao
Jiangsu Yanjiang Institute of Agricultural Sciences, Nantong 226541, China

soy flour, soy protein isolates and soy protein concentrates (Singh et al. 2008). A high level of solubility is essential for the functional characteristics of these protein forms. In addition, the utility of soybean proteins in foodstuffs of moderate acidity may be reduced when solubility is low, particularly when the desired functional properties are directly linked to solubility (Walstra 1989; Lu et al. 2013). Therefore, solubility is very important for the application of soybean protein product, and high solubility is normally a desired functional property. Although elevated solubility is not a requirement for all functional properties of food proteins, the percentage of extractable soluble protein is a traditional criterion in the evaluation of soybean product quality (Malhotra and Coupland 2004; Walsh et al. 2003).

In soybean, protein content (PC) and water-soluble protein content (SPC) are complex quantitative traits involving multiple genes. Although plant breeding through phenotypic selection has resulted in major progress in enhancing PC and SPC, this approach is time-consuming and laborious. Increasing the genotype selection intensity by marker-assisted selection (MAS) would improve the development of cultivars with increased PC and SPC. However, most QTL/gene studies have focused on PC. More than 100 QTL related to soybean PC have been reported over the past decade (http://www.soybase.org/), but only a few QTL underlying SPC have been reported (Lu et al., 2013). Moreover, these QTL encompass a relatively larger genomic region because linkage mapping is limited by allelic diversity and QTL resolution (Bergelson and Roux 2010; Zhu et al. 2008). Genome-wide association studies (GWASs) based on linkage disequilibrium (LD) overcome these limitations and have recently been successfully applied to map QTL and genes in plants, including *Arabidopsis* (Atwell et al. 2010), rice (Cai et al. 2013; Huang et al. 2011), maize (Kump et al. 2011), soybean (Sonah et al. 2013) and other plants (Cockram et al. 2010; Zhu et al. 2008). Association mapping is a powerful technique that is used to study genetic loci involved in the inheritance of complex traits. This technique can provide increased accuracy for the localization of QTL because of the higher recombination rate between markers and QTL alleles in random-mating populations (Flint-Garcia et al. 2005). In addition, haplotype association is likely to be more powerful than association based on single markers (Garner and Slatkin 2003; Bagos 2011). The use of haplotypes for QTL mapping could compensate for the bi-allelic limitation of SNPs and substantially improve the efficiency of QTL mapping (Lu et al. 2012; Hao et al. 2012b). In addition, haplotype–trait association analyses are helpful for the precise mapping of important genomic regions and localization of favorable alleles or haplotypes for breeding (Barrero et al. 2011; Hao et al. 2012a).

This study aimed to genetically dissect the mechanisms underlying PC and SPC in a diverse collection of soybean accessions using association mapping. The collection was chosen to represent a wide range of soybean genetic diversity and was previously successfully applied in a whole-genome association analysis of several traits (Hao et al. 2012b; Wang et al. 2011; Hu et al. 2013; Zhang et al. 2014). The objectives of the current study were: (1) to dissect genetic variation in PC and SPC in a diverse soybean panel; (2) to identify SNPs and haplotypes associated with soybean PC and SPC; and (3) to identify major loci affecting PC and SPC in soybean to enable the genetic improvement of PC and SPC. Our results revealed several major loci that significantly contribute to PC and SPC; the SPC-specific loci might be responsible for the low SPC of soybean varieties with high PC. Many QTL detected for both traits by GWAS were found to be located in or near regions where QTL for protein-related traits have been mapped by linkage analysis, suggesting that GWAS is an alternative, powerful mapping approach for identifying QTL underlying PC and SPC in soybean.

Materials and methods

Plant materials and field experiments

The association panel used in this study consisted of 192 soybean accessions (Table S1) originating from 26 different provinces and six ecological regions of soybean cultivation in China (latitude 53 to 24°N and longitude 134 to 97°E) (Wang and Gai 2002). This global collection was initially examined by Wang et al. (2011) using an association mapping approach to map resistance genes with microsatellite markers. Hao et al. (2012a) used this collection to study yield QTL employing SNP markers and haplotypes. The same population was also used in a whole-genome scan using SNP markers to identify QTL associated with seed shape traits and phosphorus efficiency-related traits (Hu et al. 2013; Zhang et al. 2014). This association panel was selected for use in this study, because this collection is representative of the diverse genetic variation in soybean PC and SPC.

The field experiments were performed in 2009, 2011 and 2012 at the following three locations: Jiangpu Experimental Station of Nanjing Agricultural University (32.1°N 118.4°E), Nanjing, in 2009 (designated as E1); Maozhuang Experimental Station of Henan Agricultural University (34.8°N 113.6°E), Zhengzhou, in 2009 (designated as E2) and 2011 (designated as E3); and the Experimental Farm of Henan Agricultural University (33.2°N 112.9°E), Fangcheng, in 2012 (designated as E4). A randomized complete-block design was used for all trials. At E1, all accessions were planted with two replications. At all other locations (E2, E3 and E4), all accessions were planted with three replications. In all four environments, each accession was planted in three rows per plot, each row was 200 cm long, and the row spacing was 50 cm.

Measurement of soybean PC and SPC

Three traits, PC, SPC and oil content, were evaluated in all environments. The PC and SPC in soybean seed were analyzed with a near-infrared spectrophotometer (NIR) seed analyzer (DA7200, Perten Instrument, Huddinge, Sweden). The calibrations were performed by Perten Instruments and the Inspection and Testing Center for Quality of Cereals and Their Products of the Ministry of Agriculture. These calibrations involved more than 700 uniform soybean samples that varied in seed PC, oil content and 146 soybean samples in seed SPC. Each sample (60 g) was fitted in a 75-mm-diameter cup that rotated during NIRS scanning. Three scans were conducted on each sample, and the data were read three times per sample and the averages were used in statistical analysis.

Genotyping and haplotype construction

The 192 soybean accessions used in this study were genotyped for 1,536 good-quality single nucleotide polymorphism (SNP) markers using the Illumina Bead lab system at the National Engineering Center for Biochip (Shanghai, China) by Hao et al. (2012a). In addition, marker profiling and haplotype construction based on the 1,142 SNPs with minor allele frequencies (MAFs) >10 % have been described in detail (Hao et al. 2012a). However, in this study, SNPs with MAFs lower than 5 % were excluded. The final set of 1,298 SNPs distributed over the whole soybean genome was used to study genetic diversity, population structure, genetic relatedness and marker–marker associations in relation to genetic distance. The average spacing between markers was approximately 0.77 Mb. From the 1,298 SNPs, 232 haplotypic loci were identified, each consisting of two or more SNPs. All haplotypes (including the rare haplotypes) were used for further analyses in this study.

Phenotypic data analysis

Statistical analysis of all phenotypic data across the four environments was conducted using the software SAS version 9.0 (SAS Institute, Inc., Cary, NC). All phenotypic data were subjected to analysis of variance (ANOVA) to compare differences in the means of traits of each accession across the four environments; this analysis was conducted using PROC GLM. The linear statistical model includes the effects of genotype, environment and the environment × genotype interaction. The decomposition of variance components was evaluated using PROC VARCOMP. The broad-sense heritability (h^2) of each trait was estimated as $h^2 = V_g/(V_g + V_e/y)$, where h^2 is broad-sense heritability, V_g is genetic variance, V_e is environmental variance and y is years. The correlation coefficients between PC, SPC, SPC:PC and oil content in soybean were calculated with PROC CORR.

Association analysis

In this study, the phenotypic data for the genotypes from the association mapping panel across four environments and the marker scores for a set of 1,298 SNPs with 232 haplotypes were used to perform marker–trait association analysis. To account for the effects of the population structure of the mapping panel and genetic relatedness among panel members, various statistical models were evaluated: (1) the GLM model without considering Q and K; (2) the GLM model considering Q, in which the Q matrix was included as a cofactor in the regression model to correct population structure; and (3) the MLM model considering Q and K, which considers both population structure and kinship as cofactors. According to the quantile–quantile (Q–Q) plot from the output of TASSEL4.0 (Bradbury et al. 2007; Yu et al. 2006), the Q + K methods were appropriate for the present study. Markers were identified as significantly associated with traits by comparison with the Bonferroni threshold ($P \leq 1/1{,}298 = 7.7\text{e} - 04$, $-\log P \geq 3.11$).

Results

Genetic diversity, population structure and genetic relatedness

The genetic diversity, population structure, genetic relatedness and haplotypes for the diverse panel had been previously described for 1,142 of the 1,526 SNPs by Hao et al. (2012a). This study differed from that of Hao et al. because we selected SNPs from the initial dataset with MAFs ≥ 5 % rather than those that had MAFs ≥ 10 %. As a result, we used 156 more SNPs and 23 more haplotypes than Hao et al. (2012a). Therefore, the final set used for analysis included 1,298 SNPs and 232 haplotypes. As a result, the genetic diversity and genetic relatedness observed in our study were slightly different from those reported by Hao et al. (2012a). For example, the average genetic diversity, heterozygosity and PIC of the 1,298 SNPs were 0.387, 0.011 and 0.302, with ranges of 0.056–0.615, 0–0.234 and 0.052–0.537, respectively. The 232 haplotypes consisted of 678 alleles, and 82 % of the haplotypes consisted of two SNPs. PICs for the 232 haplotype loci ranged from 0.075 to 0.887, with an average of 0.586. Because a significantly

Table 1 Descriptive statistics, ANOVA, broad-sense heritability and percentage of phenotypic variation explained by population structure for PC and SPC in 192 soybean accessions

Trait	Environment	Mean ± SD	Min–max	G[a]	E[b]	G × E[c]	H^2[d] (%)	R^2[e] (%)
PC	E1	40.01 ± 4.15	30.27–50.03	**	**	**	91.7	2.4
	E2	44.09 ± 4.01	37.27–50.80					
	E3	45.57 ± 1.98	38.67–52.52					
	E4	41.85 ± 3.22	34.66–49.73					
SPC	E1	25.93 ± 5.02	7.39–40.52	**	**	**	84.2	19.5
	E2	27.50 ± 4.85	8.04–40.82					
	E3	29.05 ± 4.08	10.08–45.51					
	E4	24.42 ± 3.75	6.99–37.70					

** Significant at $P < 0.001$

[a] Genotype

[b] Environment

[c] Genotype × environment

[d] Broad-sense heritability

[e] Percentage of phenotypic variation explained by population structure

higher level of allelic diversity was observed for haplotypes compared to SNP markers, it was expected that haplotypes would be more powerful tools in genetic diversity analysis and gene mapping. In addition, genetic relatedness analysis showed that more than 82 % of the kinship coefficient values were <0.05, suggesting that there was no (or weak) relatedness between pair-wise soybean accessions. A K matrix of the relatedness analysis based on the 1,298 SNPs was constructed for association analysis.

Table 2 Phenotypic correlations between PC, SPC, the SPC:PC ratio and oil content based on the means of the traits in 192 soybean accessions

Trait	PC	SPC	SPC:PC
SPC	0.336**		
SPC:PC	0.113 [ns]	0.972**	
Oil content	−0.692**	−0.480**	−0.343**

ns not significant

** Significant at $P < 0.001$

Significant variation among soybean accessions with respect to PC and SPC

The means, standard deviation, range and broad-sense heritability of PC and SPC as well as the percentage of phenotypic variation explained by the population structure are shown in Table 1. The mean PC for the individual accessions in the natural population ranged from 30.3 to 52.5 %, and the maximum value for SPC was approximately seven times the minimum value (Table 1). The SPC reached 45.5 %; however, one soybean accession had an SPC of only 7.0 %. ANOVA revealed that the genotype and environmental factors were all significant at the 0.001 probability level for PC and SPC (Table 1). Overall, each trait clearly exhibited considerable natural variation among different environments and displayed very high genetic diversity. The broad-sense heritability for PC (91.7 %) indicated that PC was less affected by environmental factors than SPC (84.2 %). The GLM model was employed to infer the effect of population structure on PC and SPC. This model demonstrated that the population structure can explain 2.5 and 19.5 % of the phenotypic variation for PC and SPC, respectively (Table 1). SPC was significantly correlated with PC ($r = 0.336$ $P < 0.001$), with the SPC:PC ratio ($r = 0.973$, $P < 0.001$) and with the oil content ($r = −0.480$, $P < 0.001$) (Table 2). There was no significant correlation between PC and SPC: PC ($r = 0.113$, $P = 0.119$), suggesting that the SPC was not necessarily high in the accessions with high PC.

SNP markers and haplotypes associated with PC and SPC

Three models accounting for the population structure of the mapping panel and genetic relatedness among panel members were used in the association mapping analysis. When the GLM method (considering Q or not) was applied in the association analysis across all accessions, most of the marker–trait associations that belonged to SPC groups were detected (Bonferroni threshold $P \leq 7.7e - 04$ or –log$P \geq 3.11$). However, most of these associations were identified as false-positive results by the quantile–quantile test (in TASSEL 4.0) because the P values from the SPC groups deviated from the expected value (Fig. 1). As shown in Fig. 1, the MLM model (Q + K) was significantly better than the GLM model with respect to reducing the effect of

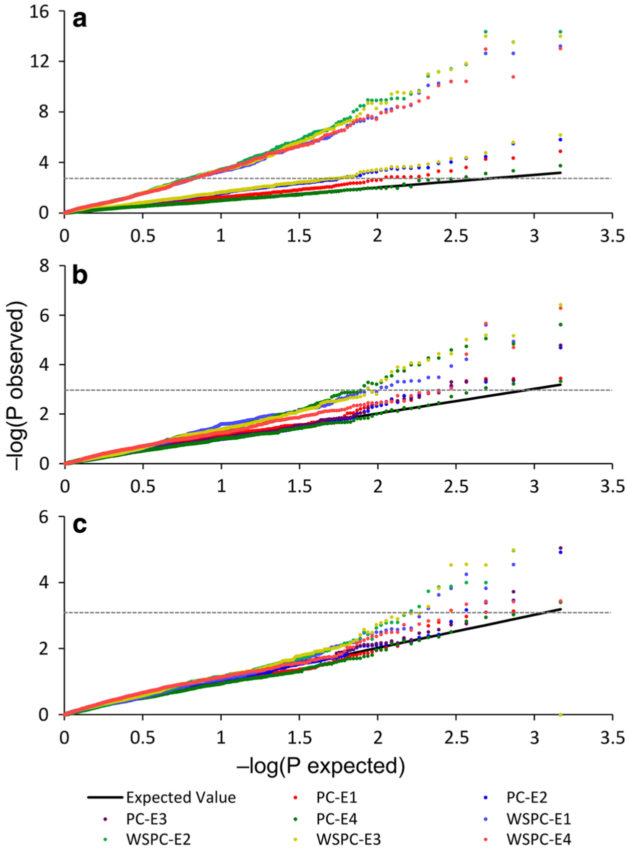

Fig. 1 Quantile–quantile (Q–Q) plots of estimated $-\log_{10}(P)$. Q–Q plots for marker–trait association analysis across four environments for PC and SPC are shown using three models: **a** the GLM method, **b** the Q method and **c** the Q + K method. The *black line* is the expected line under a null distribution. The observed *P* values are indicated in *red* for PC in E1, *blue* for PC in E2, *purple* for PC in E3, *green* for PC in E4, *light blue* for SPC in E1, *light green* for SPC in E2, *yellow* for SPC in E3 and *light red* for SPC in E4. The *gray* horizontal dashed line indicates the significance threshold of $-\log_{10}(P) \geq 3.11$

population structure and genetic relatedness on soybean PC and SPC. *P* values from the Q + K model were close to the expected values, indicating that this model is suitable for association analysis. Thus, we conducted GWAS for soybean PC and SPC with the MLM model (Q + K) to correct for population structure and genetic relatedness using 1,298 SNPs and 232 haplotypes.

As a result, 12 SNPs for PC and 24 SNPs for SPC were identified as having significant marker–trait associations at the Bonferroni-adjusted significance threshold ($P \leq 7.7\text{e} - 04$ or $-\log P \geq 3.11$) across various environments (Fig. 2). Among these significant SNPs, some were detected only in a specific environment, and only the SNPs that were identified in two or more environments are listed in Table 3. For PC, two SNPs were identified in two environments, one was identified in three environments, and one was identified in all four environments (Fig. 2; Table 3). The most significant SNP associated with PC, BARC-021267-04016 ($-\log P = 3.41$–5.04), was detected across all four environments on chromosome 19 and explained 6.8–10.9 % of the phenotypic variation. For SPC, eight SNPs were identified in two or more environments, four were identified in two environments, one was identified in three environments, and three were identified in all four environments (Fig. 2; Table 3). Compared with the QTL for PC, the most significant SNPs were detected for SPC specifically, and only one SNP was identified that was synchronously associated with both PC and SPC (Table 3). This co-associated SNP, BARC-042857-08439, was identified on chromosome 15 in environments E2 and E3 (Table 3). Other major stable QTL, such as BARC-029149-06088 on chromosome 5 as well as BARC-018023-02499 and BARC-030485-06876 on chromosome 8, were SPC-specific, suggesting that these loci may be responsible for the genetic variation in SPC between genotypes.

Haplotype mapping demonstrated that three and six haplotype loci were significantly associated with PC and SPC, respectively (Table 4). Similar to the results with SNPs, only one haplotype locus (hp191) was co-associated with PC and SPC, and most of the haplotypes were SPC-specific in two or more environments (Table 4). For PC, one haplotype (hp74) was identified across all four environments and explained 8.98–10.1 % of the phenotypic variation. A total of six significant haplotype loci were identified for SPC across two or more environments. Among them, the hp15 haplotype on chromosome 8, consisting of the three closely linked markers BARC-018023-02499, BARC-028361-05839 and BARC-028361-05840, was associated with SPC in all four environments and explained 15.1 % of the phenotypic variation. This QTL was previously identified by Lu et al. (2013) as associated with SPC in soybean. In addition, the haplotype hp131 on chromosome 18 was SPC-specific in all four environments and explained 8.2–10.0 % of the phenotypic variation.

Comparison of significant loci identified by single SNP and haplotype-based mapping

Using haplotype mapping, three and six significant haplotype loci were identified for PC and SPC across the various environments, respectively. For the nine haplotype loci identified in this study for PC and SPC, five contained at least one SNP that was also identified by single SNP-based mapping (Tables 3, 4). Furthermore, a comparison of the significant loci identified by SNP and haplotype mapping revealed that the haplotype loci tended to explain much higher proportions of phenotypic variation compared with single SNPs. For example, haplotype locus 2 (hp2), which was associated with SPC, contained the SNP BARC-029149-06088, which was also significantly associated

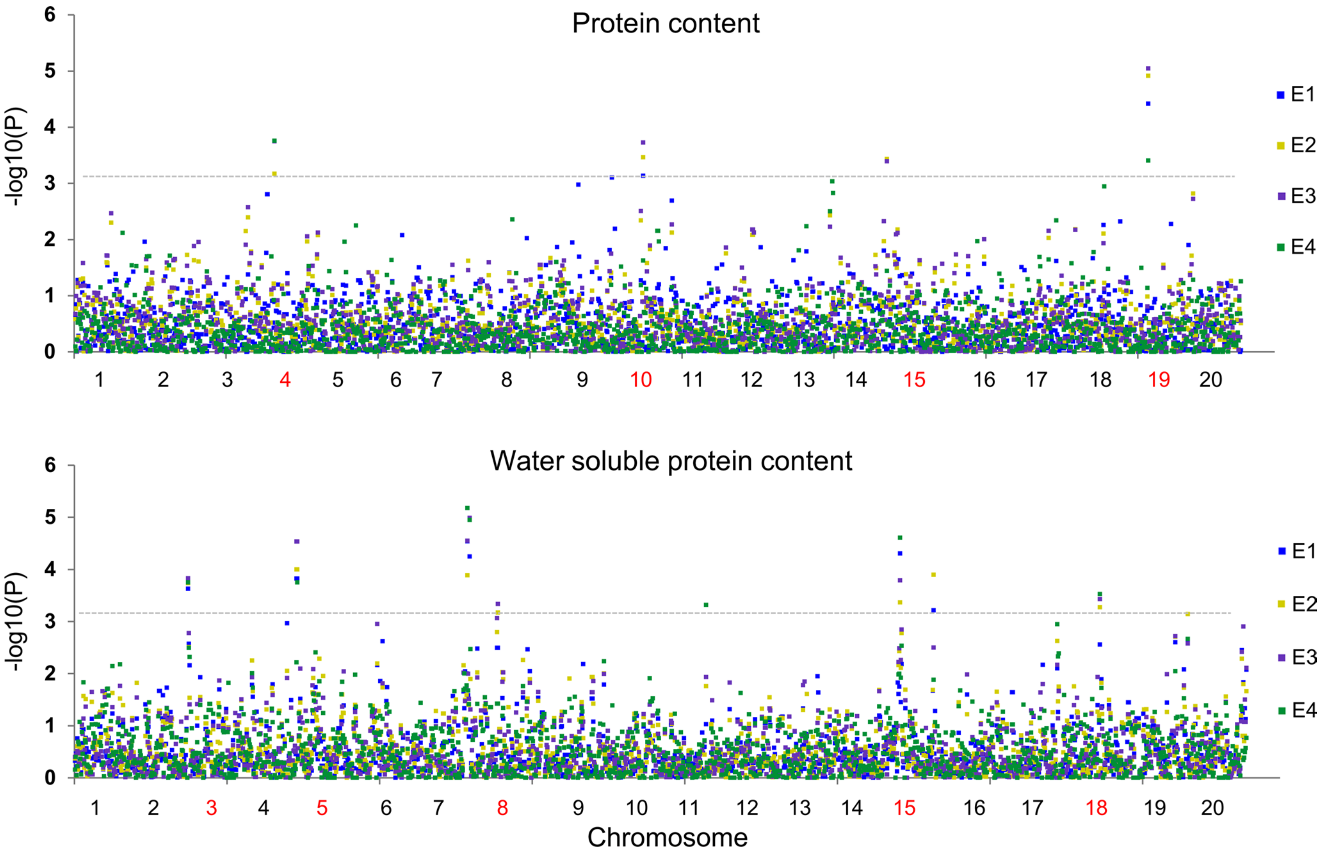

Fig. 2 Genome-wide association mapping of PC and SPC based on SNPs in four different environments (E1–E4). Chromosomes marked with *red* font denote the significant SNP clusters associated with PC or SPC that were identified in two or more environments. The *dashed line* indicates a significant association signal ($-\log P \geq 3.11$)

with SPC based on single SNP-based mapping. However, hp2 explained a much higher proportion of the phenotypic variation (10.9 %) than a single SNP (8.5 %). Within haplotype locus 15 (hp15), the SNP BARC-018023-02499 was also significantly associated with SPC and was identified by single SNP-based mapping, which explained a relatively low proportion of the phenotypic variation (10.0 %). However, its corresponding haplotype locus, which included this SNP plus two additional nearby SNPs, explained a higher proportion of the phenotypic variation (15.1 %). Within hp131 on chromosome 18, BARC-030691-06926 was also significantly associated with SPC, was identified by single SNP-based mapping and explained a relatively low proportion of the phenotypic variation (8.5 %). Its corresponding haplotype locus, which included this SNP plus an additional five nearby SNPs, could explain 13.1 % of the phenotypic variation. In conclusion, some chromosomal regions defined by haplotype loci were closely linked to the specific traits for which significant single SNP markers were identified. Furthermore, significant haplotype loci explained a much higher proportion of the phenotypic variation than single SNPs.

Allelic effects of the associated SNPs on PC and SPC

Based on marker polymorphisms, the distribution of PC and SPC was examined in 192 soybean accessions based on the four stable markers BARC-021267-04016, BARC-030485-06876, BARC-029149-06088 and BARC-018023-02499 (Fig. 3). The significant SNP BARC-021267-04016 was associated with PC across all four environments and was identified in 177 C-type and 15 G-type soybean accessions (Fig. 3a). The PC of G-type soybean accessions was significantly higher than that of C-type soybean accessions (t test, $P = 8.86 \times 10^{-13}$), with an average increase of 3.9 % per G allele. In addition, many QTL containing this SNP were previously associated with both PC and protein component content by linkage mapping using segregating populations (Orf et al. 1999; Panthee et al. 2004, 2006). For the significant SNP BARC-030485-06876, a comparison of the SPC of 34 C-type and 150 G-type soybean accessions demonstrated that the G-type accessions have a higher SPC than the C-type accessions (t test, $P = 3.12 \times 10^{-6}$) (Fig. 3b). A comparison of the SPC of 142 C-type and 43 G-type soybean accessions genotyped at

Table 3 SNPs with significant association signals ($P \leq 7.7e - 04$ or $–\log P \geq 3.11$) for soybean PC and SPC detected in two or more environments

Trait	Marker	Chr.	Position	$-\log P$[a]				MAF[c]	Related QTL[d]	References
				E1	E2	E3	E4			
PC	BARC-021803-04214	4	42715649	ns[b]	3.17	3.75	ns	0.167	*Prot 7-2*	Orf et al. (1999)
	BARC-032467-08977	10	50073557	3.14	3.46	3.73	ns	0.276	/	
	BARC-042857-08439	15	3828443	ns	3.43	3.4	ns	0.432	*Prot 17-3*	Tajuddin et al. (2003)
	BARC-021267-04016	19	42262311	3.42	4.91	5.04	3.41	0.078	*Prot 8-1* *Basic fraction 1-1* *Prot 30-7* *Prot 17-6* *Acidic fraction 1-4* *sd-Glu 1-5* *sd-Leu 1-5*	Orf et al. (1999) and Panthee et al. (2004, 2006)
SPC	BARC-028709-05992	3	31006419	3.22	3.87	ns	ns	0.221	*Beta conglycinin 1-1*	Panthee et al. (2004)
	BARC-029149-06088	5	1035515	3.83	4	4.54	3.75	0.234	/	
	BARC-018023-02499	8	7955046	4.52	3.89	4.55	5.18	0.214	*Alpha conglycinin 1-1* *sd-Ile 1-1* *sd-Glu 1-1* *sd-Asp 1-1* *qsp8-4, qsp8-5*	Panthee et al. (2004, 2006) and Lu et al. (2012)
	BARC-041663-08059	8	9622940	4.25	4.97	5	4.95	0.195	/	
	BARC-031037-06989	8	14358588	ns	3.12	3.33	ns	0.167	*Prot 26-1*	Reinprecht et al. (2006)
	BARC-042857-08439	15	3828443	ns	3.36	3.78	ns	0.173	*Prot 17-3*	Tajuddin et al. (2003)
	BARC-017679-03103	15	38000598	3.22	3.87	ns	ns	0.271	/	
	BARC-030691-06926	18	34177951	ns	3.28	3.34	3.53	0.143	*sd-Leu 1-4* *sd-Trp 1-5* *sd-Pro 1-3* *Prot 20-1*	Panthee et al. (2005, 2006)

[a] Significant at $P \leq 7.7e - 04$ or $–\log P \geq 3.11$

[b] Marker was not detected at a significant level in the corresponding environment

[c] Minor allele frequencies

[d] Previously reported protein-related QTL in SoyBase (http://www.soybase.org/)

BARC-029149-06088 demonstrated that the SPC of G-type accessions was increased by 6.7 % relative to that of the C-type accessions (t test, $P = 1.22 \times 10^{-8}$) (Fig. 3c). For the significant SNP BARC-018023-02499, which was associated with SPC, G-type accessions (36 accessions) showed higher SPC than C-type accessions (146 accessions) (Fig. 3d). The SPC increased by 8.9 % per G allele, and the SNP explained 10.1 % of the total phenotypic variation. These results confirmed that these loci might be responsible for the genetic variation in PC and SPC between genotypes and also suggest that some loci were specifically associated with SPC.

Discussion

Protein solubility is a critical factor in the food quality and yield of soybean products. Low solubility will not only result in a low yield of the food products, but also cause proteins to separate and settle out, resulting in irregular and reduced dispersion in the food products. A high level of protein solubility is required to obtain preferable emulsifying and foaming properties (Lu et al. 2013). Past efforts to improve the SPC of soybean have included enzymatic and other chemical approaches. Among these, the use of anionic surfactants, bromelain digestion and hydrolysates were effective in solubilizing soybean seed proteins (Malhotra and Coupland 2004; Molina Ortiz and Wagner 2002; Walsh et al. 2003). Although the processing of soybean protein in this manner is effective, it increases the cost of the finished product and is not a sustainable approach. Increasing the genotype selection intensity by marker-assisted selection (MAS) would facilitate the development of cultivars with enhanced PC and SPC. Unfortunately, soybean SPC is a complex trait involving multiple genes, and MAS for SPC in soybean has suffered from limitations, including issues related to resolution and genetic diversity.

Table 4 Haplotypes with significant association signals ($P \leq 7.7\mathrm{e}-04$ or $-\log P \geq 3.11$) for soybean PC and SPC detected in two or more environments

Trait	Haplotype	SNP	Chr.	Position	$-\log P$[a]			
					E1	E2	E3	E4
PC	hp74	BARC-018835-03260	1	53,570,643	3.19	3.46	3.77	4.21
		BARC-025941-05177	1	53,622,732				
	hp90	BARC-021403-04096	2	50,936,699	ns[b]	5.18	5.34	3.46
		BARC-039653-07533	2	51,233,372				
		BARC-019805-04378	2	51,243,058				
		BARC-041469-08004	2	51,406,882				
	hp191	BARC-028709-05992	3	31,006,419	ns	3.13	3.15	ns
		BARC-022187-04294	3	31,584,252				
		BARC-016467-02618	3	33,395,967				
SPC	hp191	BARC-028709-05992	3	31,006,419	3.37	3.27	3.57	4.11
		BARC-022187-04294	3	31,584,252				
		BARC-016467-02618	3	33,395,967				
	hp197	BARC-016627-02152	3	41,779,517	ns	3.12	ns	3.12
		BARC-013865-01261	3	41,779,568				
	hp2	BARC-044271-08652	5	1,023,208	3.83	4	4.54	ns
		BARC-029149-06088	5	1,035,515				
	hp15	BARC-028361-05839	8	7,716,379	4.61	3.29	3.95	5.99
		BARC-028361-05840	8	7,716,379				
		BARC-018023-02499	8	7,955,046				
	hp21	BARC-031037-06990	8	14,358,588	ns	3.11	3.22	ns
		BARC-031037-06989	8	14,358,588				
		BARC-043207-08554	8	14,435,034				
	hp131	BARC-013825-01250	18	30,313,574	3.86	3.65	4.28	3.3
		BARC-030691-06926	18	34,177,951				

[a] Significant at $P \leq 7.7\mathrm{e}-04$ or $-\log P \geq 3.11$

[b] Haplotype was not detected at a significant level in the corresponding environment

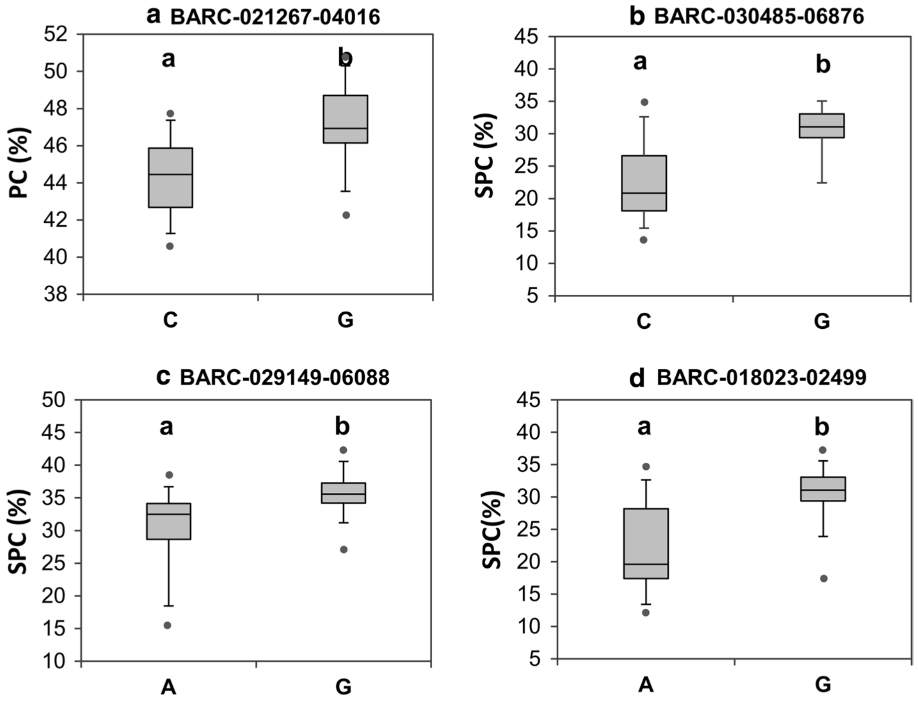

Fig. 3 Association of marker allele polymorphisms with PC and SPC. **a** *Box plot* of PC in 177 C-type and 15 G-type soybean accessions. The *vertical axis* indicates the PC. The PC of G-type accessions was significantly higher than that of C-type accessions (*t* test, $P = 8.86 \times 10^{-13}$). **b** *Box plot* of SPC in 34 C-type and 150 G-type soybean accessions. G-type accessions had significantly higher SPC (*t* test, $P = 3.12 \times 10^{-6}$). **c** The SPC of 142 C-type and 43 G-type soybean accessions; G-type accessions had higher SPC than C-type accessions (*t* test, $P = 1.22 \times 10^{-8}$). **d** Comparison of the SPC between the 146 C-type and 36 G-type accessions in the diverse panel, showing that G-type accessions have higher SPC (*t* test, $P = 2.13 \times 10^{-15}$)

In the diverse panel examined in this study, the range of PC was 30.3–52.5 %, and the range of SPC was 7.0–45.5 %, indicating greater genetic variation in PC and SPC in this study compared with previous studies (Lu et al. 2013; Pathan et al. 2013; Jun et al. 2008). Association mapping based on the performance of diverse germplasms would provide more relevant markers in a broad genetic background and enable breeders to search for favorable alleles. Heritability for PC and SPC in our study was moderately high and similar to that reported by Hyten et al. (2004) and Panthee et al. (2005). The heritability observed in our panel indicated that much of the variation was genetic; therefore, a selection response could achieve genetic gain. In addition, we observed only a relatively low correlation ($r = 0.336$, $P < 0.001$) between PC and SPC. Therefore, even though water-soluble protein is a significant component of total protein, the manner in which SPC is genetically controlled differs from the processes that affect the total PC, as reported by Lu et al. (2013).

First, given the significant correlation between soybean PC and SPC, we predicted that QTL/genes associated with PC and SPC may have pleiotropic effects. However, the reverse was also true: several PC and SPC QTL were identified independently in our study and mapped to different genomic regions. No QTL have been reported with pleiotropic effects that simultaneously increase both PC and SPC, and there has been no comparison of the genetic loci that control the two traits. In this study, we compared association mapping results for PC and SPC with QTL previously reported within a 2-Mb vicinity. The Williams 82 physical map and the soybean whole-genome sequence in SoyBase (http://www.soybase.org/pmd/index.php) were used for comparison and discussion of sequence-based genetic markers, comparative analyses with other genomes and various informatic analyses. The significant SNP BARC-042857-08439 and the haplotype hp191 were consistently associated with PC and SPC across various environments (Tables 3, 4). The SNP marker BARC-042857-08439 on chromosome 15, which was associated with PC and SPC, was located in the region of the PC QTL Prot 4-5 (Lee et al. 1996), Prot 17-3, Prot 30-3 (Tajuddin et al. 2003) and Prot 31-2 (Pandurangan et al. 2012). The significant chromosomal region hp191 on chromosome 3 is a novel QTL that has not been previously identified. These results indicate that some causal gene/genes might exist in these regions, and these associated markers may be useful for the aggregation of causal genes of interest to improve soybean PC and SPC simultaneously. In addition, we reconfirmed previous findings indicating that protein QTL flanked by SNP markers and haplotypes may be important targets that could lead to the identification of candidate genes involved in the modification of protein levels using genetic and genomic approaches.

However, the relatively low correlation between PC and SPC suggests that the manner in which SPC is genetically controlled differs from the processes that affect PC. The results also demonstrated that most QTL detected in our study were independent and mapped to different genomic regions (Tables 3). For PC, BARC-021267-04016 on chromosome 19 was located in the same region as QTL for seed protein content (Hyten et al. 2004; Orf et al. 1999; Panthee et al. 2006), protein components (Panthee et al. 2004) and the filling rate of protein (Jiang et al. 2010). Moreover, this locus was localized 0.28 cM away from a published glycinin gene, G7, a major gene controlling the synthesis and assembly of glycinin in soybean (Beilinson et al. 2002). Another significant SNP, BARC-021803-04214, was also associated with PC by Orf et al. (1999). On chromosome 8, the loci containing BARC-018023-02499, BARC-041663-08059 and hp15 were identified as significantly associated with SPC by SNP and haplotype-based mapping in our study. Previously, two key QTL (*qsp8-4* and *qsp8-5*) related to SPC were identified in this region by Lu et al. (2013) using linkage mapping with an RIL population. QTL for α-conglycinin and PC were previously mapped in the same genomic region (Panthee et al. 2006; Pathan et al. 2013). Moreover, we predict that a candidate gene, glutelin type-B 2-like (0.49 cM away from the marker BARC-041663-08059), which belongs to the cupin superfamily and functions in the storage of nutrient substrates, might play an important role in determining soybean protein content (Dunwell 1998; Dunwell et al. 2004). In addition, four significant SNP markers, BARC-028709-05992, BARC-041663-08059, BARC-031037-06989 and BARC-030691-06926 on chromosomes 3, 8 and 18, were determined to be associated with SPC, coinciding with a previously mapped soy SPC QTL (Lu et al. 2013), beta conglycinin (Panthee et al. 2004) and other protein-related components (Tajuddin et al. 2003; Reinprecht et al. 2006; Panthee et al. 2005; Panthee et al. 2006). These SPC-specific loci may be responsible for the genetic variation in SPC between genotypes. In this study, we also detected several novel SNPs and haplotypes that were significantly associated with soybean PC and SPC, including BARC-032467-08977 and hp74 on chromosomes 10 and 1 (associated with PC) and BARC-029149-06088, BARC-017679-03103 and hp2 on chromosomes 5, 15 and 5 (associated with SPC), which generally had considerable effects and displayed expression stability across various environments. These new loci are attractive candidate regions that may help to further elucidate the genetic basis of PC and SPC in soybean.

In conclusion, this study represents the latest analysis of the genetic basis of PC and SPC using GWAS based on single SNPs and haplotypes in soybean. The present study demonstrates the power of whole-genome association analysis to identify phenotype–genotype relationships and genomic regions underlying quantitative variation. By comparing the correlation between and genetic loci underlying PC and SPC, we provide new insights into the low SPC of soybean varieties with high PC. These SPC-specific QTL will be invaluable in breeding new varieties with high SPC. In addition, this study provides an example of the use of both single SNP markers and their combinations/haplotypes for improving association mapping. Our study also lays the foundation for understanding the genetic basis of SPC, which will enable the enhancement of SPC in soybean. The favorable alleles and haplotypes not only facilitate more efficient selection of soybean genotypes with high levels of SPC, but also indicate that this method can be used to enhance the power of QTL mapping for other quantitative traits in soybean.

Acknowledgments This work was supported by the National Basic Research Program of China (973 Program) (2010CB125906), the National Natural Science Foundation of China (31171573, 31201230, 31301336, 31301341) and the Jiangsu Provincial Programs (BE2012328, BK2012768).

Conflict of interest We declare that we have no conflicts of interest.

References

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465(7298):627–631

Bagos PG (2011) Meta-analysis of haplotype-association studies: comparison of methods and empirical evaluation of the literature. BMC Genet 12(1):8

Barrero RA, Bellgard M, Zhang X (2011) Diverse approaches to achieving grain yield in wheat. Funct Integr Genomic 11(1):37–48

Beilinson V, Chen Z, Shoemaker R, Fischer R, Goldberg R, Nielsen N (2002) Genomic organization of glycinin genes in soybean. Theor Appl Genet 104(6–7):1132–1140

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. Nat Rev Genet 11(12):867–879

Cai S, Wu D, Jabeen Z, Huang Y, Huang Y, Zhang G (2013) Genome-wide association analysis of aluminum tolerance in cultivated and Tibetan wild barley. Plos One 8(7):e69776

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Natl Acad Sci 107(50):21611–21616

Dunwell JM (1998) Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins. Biotechnol Genet Eng Rev 15(1):1–32

Dunwell JM, Purvis A, Khuri S (2004) Cupins: the most functionally diverse protein superfamily? Phytochemistry 65(1):7–17

Erdman JW Jr (2000) Soy protein and cardiovascular disease: a statement for healthcare professionals from the nutrition committee of the AHA. Circulation 102(20):2555–2559

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44(6):1054–1064

Garner C, Slatkin M (2003) On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. Genet Epidemiol 24(1):57–67

Hao C, Wang Y, Hou J, Feuillet C, Balfourier F, Zhang X (2012a) Association mapping and haplotype analysis of a 3.1-Mb genomic region involved in fusarium head blight resistance on wheat chromosome 3BS. Plos One 7(10):e46444

Hao D, Cheng H, Yin Z, Cui S, Zhang D, Wang H, Yu D (2012b) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. Theor Appl Genet 124(3):447–458

Hu Z, Zhang H, Kan G, Ma D, Zhang D, Shi G, Hong D, Zhang G, Yu D (2013) Determination of the genetic architecture of seed size and shape via linkage and association analysis in soybean (*Glycine max* L. Merr.). Genetica 141(4–6):1–8

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44(1):32–39

Hyten D, Pantalone V, Sams C, Saxton A, Landau-Ellis D, Stefaniak T, Schmidt M (2004) Seed quality QTL in a prominent soybean population. Theor Appl Genet 109(3):552–561

Jiang Z, Han Y, Teng W, Zhang Z, Sun D, Li Y, Li W (2010) Identification of QTL underlying the filling rate of protein at different developmental stages of soybean seed. Euphytica 175(2):227–236

Jun T-H, Van K, Kim MY, Lee S-H, Walker DR (2008) Association analysis using SSR markers to find QTL for seed protein content in soybean. Euphytica 162(2):179–191

Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, McMullen MD, Ware D (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43(2):163–168

Lee S, Bailey M, Mian M, Carter T Jr, Shipe E, Ashley D, Parrott W, Hussey R, Boerma H (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor Appl Genet 93(5–6):649–657

Lu Y, Xu J, Yuan Z, Hao Z, Xie C, Li X, Shah T, Lan H, Zhang S, Rong T (2012) Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. Mol Breed 30(1):407–418

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(2):425–433

Malhotra A, Coupland JN (2004) The effect of surfactants on the solubility, zeta potential, and viscosity of soy protein isolates. Food Hydrocoll 18(1):101–108

Molina Ortiz SE, Wagner JR (2002) Hydrolysates of native and modified soy protein isolates: structural characteristics, solubility and foaming properties. Food Res Int 35(6):511–518

Orf J, Chase K, Jarvik T, Mansur L, Cregan P, Adler F, Lark K (1999) Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Sci 39(6):1642–1651

Pandurangan S, Pajak A, Molnar SJ, Cober ER, Dhaubhadel S, Hernández-Sebastià C, Kaiser WM, Nelson RL, Huber SC, Marsolais F (2012) Relationship between asparagine metabolism and protein concentration in soybean seed. J Exp Bot 63(8):3173–3184

Panthee D, Kwanyuen P, Sams C, West D, Saxton A, Pantalone V (2004) Quantitative trait loci for β-conglycinin (7S) and glycinin

(11S) fractions of soybean storage protein. J Am Oil Chem Soc 81(11):1005–1012

Panthee D, Pantalone V, West D, Saxton A, Sams C (2005) Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci 45:2015–2022

Panthee D, Pantalone V, Saxton A, West D, Sams C (2006) Genomic regions associated with amino acid composition in soybean. Mol Breed 17(1):79–89

Pathan SM, Vuong T, Clark K, Lee J-D, Shannon JG, Roberts CA, Ellersieck MR, Burton JW, Cregan PB, Hyten DL (2013) Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. Crop Sci 53(3):765–774

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

Singh P, Kumar R, Sabapathy SN, Bawa AS (2008) Functional and edible uses of soy protein products. Compr Rev Food Sci Food Saf 7(1):14–28. doi:10.1111/j.1541-4337.2007.00025.x

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. Plos One 8(1):e54603

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(2):133–140

Walsh DJ, Cleary D, McCarthy E, Murphy S, FitzGerald RJ (2003) Modification of the nitrogen solubility properties of soy protein isolate following proteolysis and transglutaminase cross-linking. Food Res Int 36(7):677–683

Walstra P (1989) Principles of foam formation and stability. In: Wilson AJ (ed) Springer, London pp 1–15

Wang Y, Gai J (2002) Study on the ecological regions of soybean in China. II. Ecological environment and representative varieties. J Appl Ecol 13(1):71–75

Wang H, Gao Z, Zhang D, Cheng H, Yu D (2011) Identification of genes with soybean resistance to common cutworm by association analysis. Chin Bull Bot 46(5):514–524

Yu J, Pressoir G, Briggs W, Vroh Bi I, Yamasaki M, Doebley J, McMullen M, Gaut B, Nielsen D, Holland J (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203–208

Zhang D, Song H, Cheng H, Hao D, Wang H, Kan G, Jin H, Yu D (2014) The acid phosphatase-encoding gene gmacp1 contributes to soybean tolerance to low-phosphorus stress. Plos Genet 10(1):e1004061

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1(1):5–20