Genomics 111 (2019) 90–95

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions

Dongmei Li[a], Xue Zhao[b], Yingpeng Han[b], Wenbin Li[b,*], Futi Xie[a,*]

[a] Shenyang Agricultural University, Soybean Research Institute, Shenyang 110866, Liaoning, China
[b] Northeast Agricultural University, Northeastern Key Lab Soybean Biol & Genet & Breed, Chinese Ministry of Agriculture, Key Lab Soybean Biology, Chinese Ministry of Education, Harbin 150030, Heilongjiang, China

ARTICLE INFO

Keywords:
Seed protein content
Seed oil content
Soybean
Genome-wide association

ABSTRACT

Soybean is globally cultivated primarily for its protein and oil. The protein and oil contents of the seeds are quantitatively inherited traits determined by the interaction of numerous genes. In order to gain a better understanding of the molecular foundation of soybean protein and oil content for the marker-assisted selection (MAS) of high quality traits, a population of 185 soybean germplasms was evaluated to identify the quantitative trait loci (QTLs) associated with the seed protein and oil contents. Using specific length amplified fragment sequencing (SLAF-seq) technology, a total of 12,072 single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) ≥ 0.05 were detected across the 20 chromosomes (Chr), with a marker density of 78.7 kbp. A total of 31 SNPs located on 12 of the 20 soybean chromosomes were correlated with seed protein and oil content. Of the 31 SNPs that were associated with the two target traits, 31 beneficial alleles were identified. Two SNP markers, namely rs15774585 and rs15783346 on Chr 07, were determined to be related to seed oil content both in 2015 and 2016. Three SNP markers, rs53140888 on Chr 01, rs19485676 on Chr 13, and rs24787338 on Chr 20 were correlated with seed protein content both in 2015 and 2016. These beneficial alleles may potentially contribute towards the MAS of favorable soybean protein and oil characteristics.

1. Introduction

Soybean (*Glycine max*) is an economically important crop that is valued for its vegetable oil and protein. The oil and protein present in the seed constitute essential quality traits in breeding programs. A diversity of soybean cultivars from across the world have previously been used for evaluating and improving seed oil and protein content. Comparisons between landraces and modern cultivars has revealed that in the past, breeders focused on improving oil content [1]. Although these two traits can be improved via traditional breeding methods, the development of soybean lines with high contents of both is challenging, as oil and protein are negatively correlated [2]. Marker-assisted selection (MAS) is a far more efficient means of achieving this [3].

QTLs mapping is a useful approach for dissecting complex traits at the molecular genetic level in plants. Several QTLs/genes that are influenced by the environment control protein and oil contents in seeds. Numerous studies have reported on the QTLs associated with protein and oil contents in seeds (http://www.soybase.org/). These QTLs were detected via linkage analysis of populations derived from the crosses of two parents exhibiting contrasting seed protein and oil concentrations, and have been discovered in various genomic regions across all 20 chromosomes [4–8]. However, this method is hampered by its comparatively low genomic resolution when assessing the recombination events within the mapping populations, and is only able to narrowly capture the allelic diversity present in the two parental lines. Conversely, genome-wide association studies (GWAS) can utilize historic recombination events within natural populations, thereby overcoming the limitations of QTL mapping [9]. GWAS provides comparatively higher resolution with respect to determining the genomic position of a gene or QTL, as the collections of unrelated genotypes exhibit far more limited linkage disequilibrium between pairs of neighboring markers in the GWAS approach [10].

Sequencing costs have been radically reduced due to the development of high-throughput sequencing technology. This is particularly

Abbreviations: MAS, marker-assisted selection; QTLs, quantitative trait loci; MAF, minor allele frequency; SLAF-seq, specific length amplified fragment sequencing; SNPs, single nucleotide polymorphisms; Chr, chromosomes; PCA, principal component analysis; GWAS, genome-wide association studies; LD, linkage disequilibrium; NGS, next-generation sequencing technology; RAD-seq, restriction site-associated DNA sequencing; 2b-RAD, 2b-restriction site-associated DNA; GBS, genotyping-by-sequencing; MLM, mixed linear model; SSRs, simple sequences repeats; RAPD, random amplified polymorphic DNA; RFLPs, restriction fragment length polymorphisms

* Corresponding authors.
E-mail addresses: wenbinli@neau.edu.cn (W. Li), snsoybean@sohu.com (F. Xie).

https://doi.org/10.1016/j.ygeno.2018.01.004
Received 15 September 2017; Received in revised form 4 December 2017; Accepted 7 January 2018
Available online 08 January 2018
0888-7543/ © 2018 Published by Elsevier Inc.

Table 1
Analysis of variance of seed protein and oil content.

Traits	Environments	Mean ± SE (%)	Range (%)	CV (%)	Skewness	Kurtosis	Genotype	Environment
Oil	2015	20.43 ± 0.08	16.75–23.11	1.03	−0.79	1.00	21.45**	301.23**
	2016	21.88 ± 0.10	16.8–24.8	1.36	−0.92	1.30		
Protein	2015	41.64 ± 0.14	29.87–46.40	1.85	−1.29	8.40	2.27**	46.33**
	2016	42.39 ± 0.14	36.6–48.7	1.91	0.27	0.31		

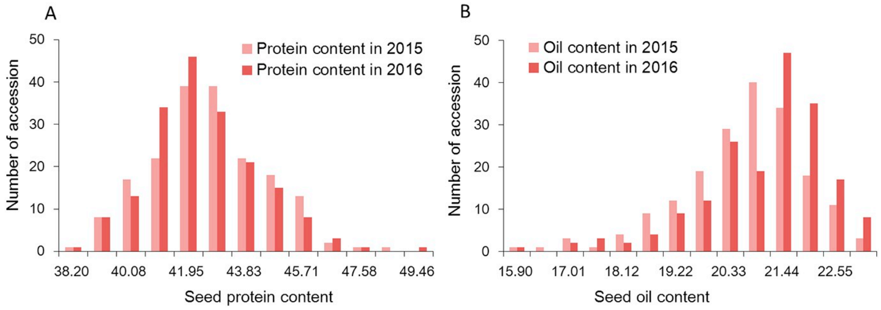

Fig. 1. Distribution of seed protein (A) and oil content (B) among 185 soybean accessions.

true for sequences of reduced representation libraries, where sequencing costs can be reduced by only sequencing representative parts of a complex genome. Several genotyping-based methods based on next-generation sequencing technology (NGS) exist, for instance restriction site-associated DNA sequencing (RAD-seq) [11], 2b-restriction site-associated DNA (2b-RAD) [12], genotyping-by-sequencing (GBS) [13], and specific length amplified fragment sequencing (SLAF-seq). SLAF-seq constitutes an intermediate between higher genotyping accuracy and relatively lower sequencing costs [14], and is therefore highly suitable for genetic association studies.

The application of GWAS has benefitted from the advance of next-generation genome sequencing technologies, and numerous GWAS have been successfully conducted in several plant species, such as *Arabidopsis* [15], rice [16], maize [17], barley [18], tomato [19], oat [20], and sorghum [21]. The loci associated with important agronomic traits, abiotic stress [22], and disease resistance have been identified by GWAS in soybean, including *Phytophthora* root rot [23], *Sclerotinia* stem rot [24, 25], soybean cyst nematode [26–28], and sudden death syndrome [29].

To investigate the genetic basis of variation in oil and protein content in soybean seeds, a diverse collection of 421 predominantly Chinese soybean cultivars was genotyped using 1536 SNPs, mostly from candidate genes related to acyl-lipid metabolism and from regions harboring known QTLs [1]. Hwang et al. [30] assessed seed protein and oil composition in soybean using a GWAS of over 55,000 SNPs across a diverse set of 298 accessions. The list of previously reported QTLs for protein and oil content was significantly reduced by their study. Sonah et al. [10] performed a GWAS for oil and protein content in a subset of 139 short-season soybean lines and included six simple morphological traits, using over 17,000 SNPs generated using GBS approaches. The authors were able to successfully identify highly significant associations for the SNPs in the candidate genes as a result of their high-resolution marker coverage.

In order to identify the QTLs related to soybean protein and oil content, we used a SLAF-seq approach for the whole-genome genotyping of a population of 185 soybean germplasms. Additionally, the genetic basis of the traits associated with high-quality protein and oil content for MAS was elucidated.

2. Materials and methods

2.1. Genotyping of soybean germplasms

A natural population consisting of 185 diverse soybean accessions was collected from 43°N to 50°N which encompassed most of the northern regions of China and other countries including America, Canada, Japan and some European countries.

These accessions were used for the phenotypic assessment of seed protein and oil content, as well as the GWAS. The genomic DNA was extracted from the leaves of each accession based on the method of Wu et al. [31] and sequenced using the SLAF-seq methodology [32, 33]. In order to obtain > 50,000 reads (SLAF-tags) per genome, different restriction enzyme combinations were tested. Enzymes were selected based on SLAF alignments to the reference genome sequence of Williams 82 (NSRL, Champaign, IL, USA) [34], and two restriction enzymes (MseI and HaeIII) were selected. Different length fragments of genomic DNA after digestion were simulated in silico. The 45-bp read at both ends of each simulated 500–550 bp fragment was sequenced on an Ilumina Genome Analyzer. The minor allele frequency (MAF) threshold was set to 0.1 in the SNP calling, and a depth of minor allele/the total

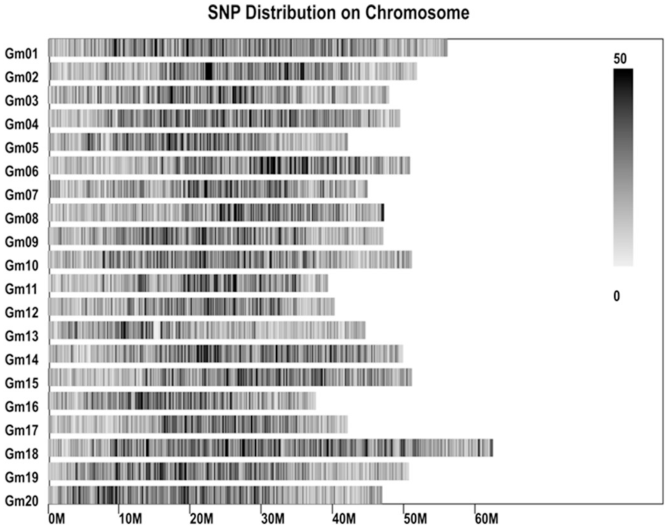

Fig. 2. SNP distribution on 20 soybean chromosomes.

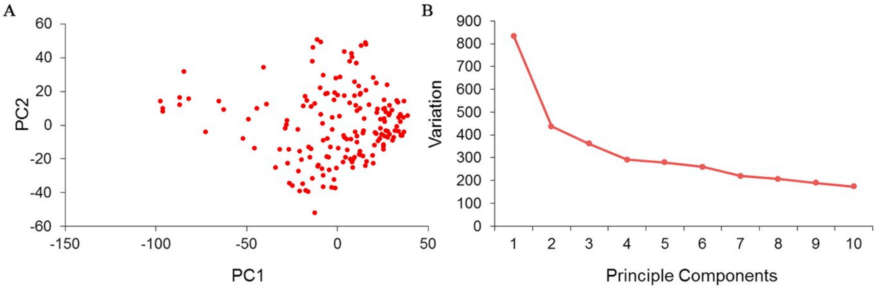

Fig. 3. PCA of population structure. (A) Distribution of the accessions in the association panel under PC1 and PC2. (B) The genetic variation explained by the first ten 10PCs.

depth of the sample ≥ 1/3 signified a heterozygous genotype.

2.2. Field trials and assessment of soybean protein and oil content

All the soybean accessions were planted in the Experimental Station of the Northeast Agricultural University in Harbin City (117°17′E, 33°18′N) with two replications in 2015 and 2016. A completely randomized block design was used for the field experiments. Each line was 3 m in length and 0.65 m apart, with 6-cm spacing between two plants. Seeds were harvested from exactly 10 plants from each plot of a single genotype, and were subsequently used in the protein and oil content determination. An Infratec 1241 NIR Grain Analyzer (FOSS, Sweden) was used to analyze three seed samples from each plot (approximately 20–25 g of seeds).

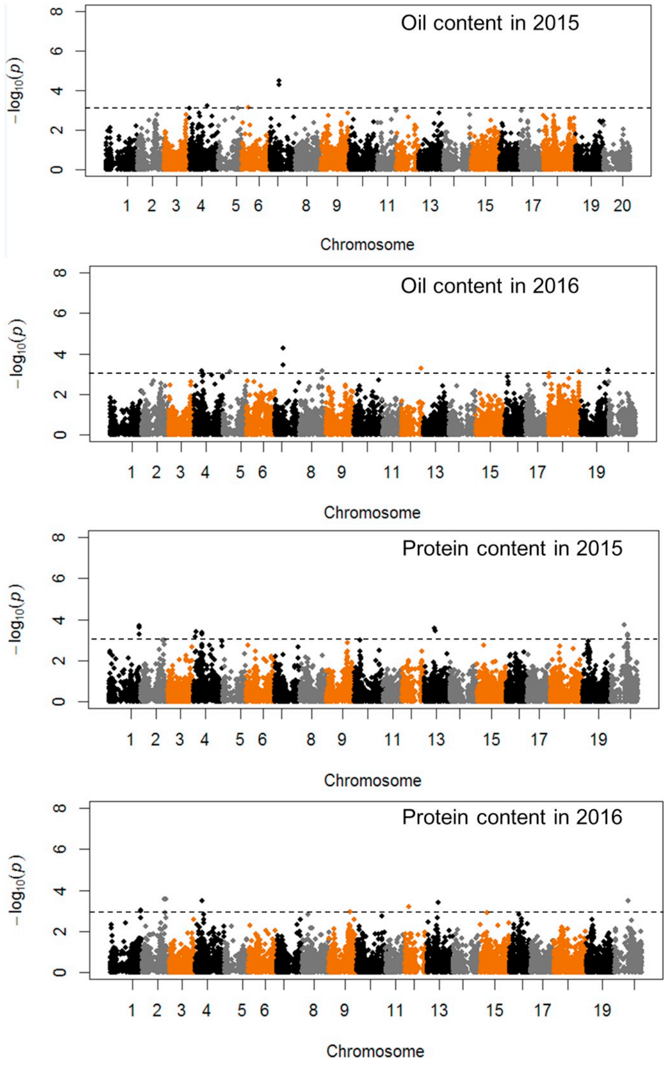

Fig. 4. Genome-wide Manhattan plots of associations for seed protein and oil content based on the compressed MLM. The *x*-axis indicates the SNPs along with each chromosome; the *y*-axis is the −log 10 (*P*-value) for the association.

2.3. Population structure evaluation

Principal component analysis (PCA) was used to assess the population structure using the GAPIT software package [35].

2.4. Association mapping

A compressed mixed linear model (MLM) in GAPIT [35] was used for the GWAS based on the SNPs from the 185 soybean accessions. A *P*-value of 0.001 constituted the Type I error significance threshold [30]. The seed protein and oil genomic QTL locations from previous studies were compared with the physical positions of the markers exhibiting significant associations in this study as a means of verifying the identified genomic regions.

3. Results

3.1. Seed protein and oil phenotyping

Protein and oil contents for the 185 soybean accessions were determined based on the dry seed weight in the 2015 and 2016 field trials. Substantial variation was observed in both traits (Table 1). The oil content ranged from 15.9–23.1% and the protein content ranged from 38.2–50.4%. A significant negative correlation was observed between seed oil and protein content, with a correlation coefficient of −0.53 ($P < .01$). The kurtosis was 0.58 and 0.92 and the skewness was 0.73 and 0.85 for protein content in 2015 and 2016, respectively. For oil content, the kurtosis was 0.99 and −0.80 and the skewness was −0.77 and 0.98 in 2015 and 2016, respectively. Following normalization, the phenotypic data of the two target traits were nearly normally distributed (Fig. 1).

SNP genotyping and population structure of the association

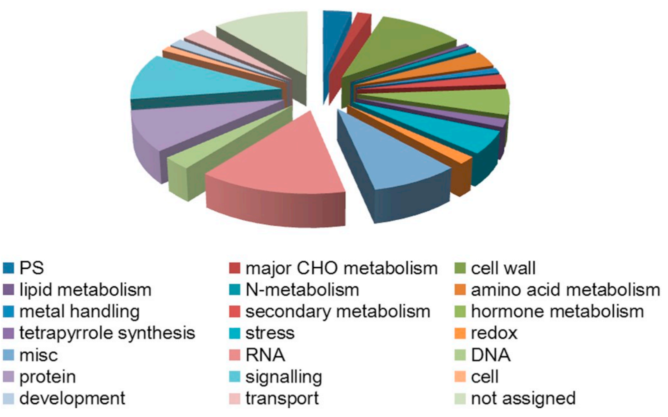

Fig. 5. Functional classification for genes in the 50 Kbp flanking region of peak SNPs.

Table 2
SNPs associated with seed protein content.

SNP	Chromosome	Year	Physical position	-log$_{10}$(P)	Minor Allele Frequency	R^2 of Model with SNP (%)	Allele 1	Allele 2	Average seed protein content of accessions with allele 1	Average seed protein content of accessions with allele 2	Average seed protein content of population
rs53140888	1	2015	53,140,888	3.68	0.26	16.03	T	C	42.80	41.54	42.00
rs53141323	1	2015	53,141,323	3.62	0.27	15.88	G	A	42.80	41.54	42.00
rs53141571	1	2015	53,141,571	3.29	0.26	15.12	T	C	42.71	41.54	42.00
rs41860441	2	2015	41,860,441	3.01	0.22	14.46	T	G	42.70	41.66	42.00
rs4187467	4	2015	4,187,467	3.15	0.17	14.79	T	A	43.04	41.64	42.00
rs4553067	4	2015	4,553,067	3.40	0.19	15.38	C	T	42.96	41.59	42.00
rs18150116	13	2015	18,150,116	3.56	0.11	15.75	T	C	43.44	41.69	42.00
rs19485676	13	2015	19,485,676	3.45	0.12	15.49	C	A	43.38	41.68	42.00
rs24787338	20	2015	24,787,338	3.74	0.13	16.17	G	T	43.54	41.64	42.00
rs29457452	20	2015	29,457,452	3.18	0.10	14.85	A	T	43.77	41.69	42.00
rs29979450	20	2015	29,979,450	3.27	0.07	15.07	T	C	43.95	41.73	42.00
rs53140888	1	2016	53,140,888	3.01	0.26	13.58	T	C	42.92	41.70	42.05
rs39587099	2	2016	39,587,099	3.55	0.20	14.85	T	A	43.23	41.73	42.05
rs41989469	2	2016	41,989,469	3.57	0.09	14.90	A	T	44.14	41.83	42.05
rs6121092	12	2016	6,121,092	3.19	0.15	14.01	A	T	42.74	41.91	42.05
rs19485676	13	2016	19,485,676	3.41	0.12	14.53	C	A	43.57	41.82	42.05
rs24787338	20	2016	24,787,338	3.50	0.13	14.72	G	T	43.80	41.77	42.05

Table 3
SNPs associated with seed oil content.

SNP	Chromosome	Year	Physical position	-log$_{10}$(P)	Minor Allele Frequency	R^2 of Model with SNP (%)	Allele 1	Allele 2	Average seed oil content of accessions with allele 1	Average seed oil content of accessions with allele 2	Average seed oil content of population
rs204699	4	2015	204,699	3.12	0.47	29.28	C	T	20.90	20.51	20.61
rs32697665	4	2015	32,697,665	3.21	0.16	29.45	A	T	20.83	19.95	20.61
rs35315668	5	2015	35,315,668	3.11	0.18	29.25	G	A	21.26	20.56	20.61
rs15774585	7	2015	15,774,585	4.48	0.06	31.96	G	T	20.76	19.62	20.61
rs15783346	7	2015	15,783,346	4.27	0.10	31.54	T	C	20.76	20.03	20.61
rs14297133	4	2016	14,297,133	3.14	0.20	33.03	G	A	20.99	20.49	20.89
rs12023370	5	2016	12,023,370	3.10	0.11	32.95	C	T	20.98	20.23	20.89
rs15774585	7	2016	15,774,585	4.26	0.06	35.11	G	T	20.97	19.78	20.89
rs15783346	7	2016	15,783,346	3.44	0.10	33.58	T	C	20.97	20.24	20.89
rs40944487	8	2016	40,944,487	3.15	0.49	33.04	G	A	21.13	20.67	20.89
rs35333667	12	2016	35,333,667	3.27	0.43	33.26	C	T	20.95	20.81	20.89
rs1452418	18	2016	1,452,418	3.04	0.31	32.85	C	A	21.19	20.77	20.89
rs55473106	18	2016	55,473,106	3.12	0.15	32.99	T	C	21.02	20.87	20.89
rs49097495	19	2016	49,097,495	3.21	0.13	33.15	G	T	20.96	20.35	20.89

mapping panel.

The genotyped samples included 185 soybean germplasms from a Chinese core collection, including elite varieties and landraces. The genomic DNA of these 185 accessions was partially sequenced using SLAF-seq and the Illumina Genome Analyzer II [32]. A total of 12,072 SNPs with an MAF ≥ 0.05 were detected across the 20 chromosomes with a marker density of 78.7 kb (Fig. 2).

Principal Component 1 (PC1) explained 7.96% of the variation in the genotypic data, while PC2 and PC3 explained 3.59% and 3.29% of the variation, respectively. The tested accessions could not be obviously grouped based on the first two axes of the PCA (Fig. 3A). However, an assessment of the variation of the first 10 axes of the PCA revealed an inflection point at PC3 (Fig. 3B). This suggested that the first three PCs dictated the impact of population structure on the association mapping, and were thus included in the compressed MLM for the association analyses.

3.2. GWAS on the loci underlying seed protein and oil content

The GWAS revealed a total of 31 SNPs associated with the two target traits, located on 12 of the 20 chromosomes (Chr.) under Compressed Mixed Linear Model (CMLM) (Figs. 2 and 3). Six and nine SNPs associated with seed oil content were respectively identified in 2015 and 2016, representing 12 genomic regions covering eight soybean chromosomes. Two SNP markers, rs15774585 and rs15783346 on Chr 07, were identified as associated with seed oil content in both 2015 and 2016. With respect to seed protein content, 14 and seven SNPs were respectively detected in 2015 and 2016, which represented 13 genomic regions covering six chromosomes. Three SNP markers, rs53140888 on Chr 01, rs19485676 on Chr 13, and rs24787338 on Chr 20 were associated with seed protein content in both 2015 and 2016. The above five SNPs were stable across the different years (Fig. 4).

In order to verify the beneficial allele (the allele that has the effect of improving the phenotype value of the target trait) of each SNP associated with seed oil and protein contents, the average seed protein and oil contents were calculated from the soybean accessions that possessed each SNP allele (Tables 2 and 3). The average seed protein content of the accessions possessing the beneficial allele (allele 1) was higher than that of the accessions with allele 2 and the entire association panel (Table 2). The difference of seed protein content between accessions with allele 1 and accessions with allele 2 reached 0.83–2.30 percentage points. The difference of seed protein content were 0.69–2.09 percentage points between accessions with allele 1 and the entire association panel. The mean seed oil content of the accessions with the beneficial allele (allele 1) was 0.14–1.20 percentage points higher than the accessions with allele 2, and was 0.06–0.65 percentage

Table 4
The overlap or linkage relationship of peak SNP and known QTLs associated with seed protein and oil content.

Trait	Year	SNP	Chromosome	Position (bp)	QTL	QTL related trait	Left marker	Right marker	Interval (bp)	Reference
Seed oil content	2015	rs35315668	5	35,315,668	Oil 4-1	Seed oil content	Sat_407	SOYNOD26A	33,718,147–35,100,953	Brummer et al. (1997)
Seed oil content	2015	rs12328685	6	12,328,685	Prot 34-2	Seed protein content	GMAC7L	Sat_213	12,232,992–14,601,295	Lu et al. (2012)
Seed oil content	2016	rs49097495	19	49,097,495	Prot 16-2	Seed protein content	BARC-032011-07238	BARC-019039-03054	48,098,859–50,424,488	Chapman et al. (2003)
Seed protein content	2015	rs4187467	4	4,187,467	Prot 12-2	Seed protein content	Sat_337	Sat_140	4,172,658–5,221,426	Specht et al. (2001)
Seed protein content	2015	rs4553067	4	4,553,067	Prot 12-2	Seed protein content	Sat_337	Sat_140	4,172,658–5,221,426	Specht et al. (2001)
Seed protein content	2016	rs6121092	12	6,121,092	Prot 28-3	Seed protein content	Sat_127	Satt442	4,265,135–6,361,515	Liang et al. (2010)
					Oil 4-10	Seed oil content	Sat_127	BARC-041917-08135	4,265,135–6,370,488	Brummer et al. (1997)
					Prot 3-11	Seed protein content	Sat_127	BARC-041917-08135	4,265,135–6,370,488	Brummer et al. (1997)
Seed protein content	2015	rs18150116	13	18,150,116	Prot 26-13	Seed protein content	Satt649	Satt325	12,953,230–18,091,080	Reinprecht et al. (2006)
Seed protein content	2015and 2016	rs19485676	13	19,485,676	Prot 26-13	Seed protein content	Satt649	Satt325	12,953,230–18,091,080	Reinprecht et al. (2006)
Seed protein content	2015 and 2016	rs24787338	20	24,787,338	Oil 2-1	Seed oil content	Satt239	BARC-027790-06672	24,129,682–32,934,647	Diers et al. (1992c)
					Prot 1-1	Seed protein content	Satt239	BARC-027790-06672	24,129,682–32,934,647	Diers et al. (1992c)
					Prot 34-11	Seed protein content	Satt700	Satt270	24,352,903–34,223,110	Lu et al. (2012)
					Oil 24-30	Seed oil content	Sat_219	Sat_105	24,528,543–34,234,025	Qi et al. (2011)
					Oil 2-2	Seed oil content	BARC-040489-07755	BARC-027790-06672	24,581,312–32,934,647	Diers et al. (1992c)
					Prot 1-2	Seed protein content	BARC-040489-07755	BARC-027790-06672	24,581,312–32,934,647	Diers et al. (1992c)
Seed protein content	2015	rs29457452	20	29,457,452	Oil 15-1	Seed oil content	Satt496	BARC-041129-07912	26,502,973–32,449,414	Chung et al. (2003)
					Prot 15-1	Seed protein content	Satt496	BARC-041129-07912	26,502,973–32,449,414	Chung et al. (2003)
Seed protein content	2015	rs29979450	20	29,979,450	Oil 15-1	Seed oil content	Satt496	BARC-041129-07912	26,502,973–32,449,414	Chung et al. (2003)
					Prot 15-1	Seed protein content	Satt496	BARC-041129-07912	26,502,973–32,449,414	Chung et al. (2003)

points higher than the entire association panel (Table 3). Based on these results, it was concluded that these beneficial alleles could be utilized in MAS for seed protein and oil traits in soybean. Moreover, the Enhanced Compressed Mixed Linear Model (ECMLM) in GAPIT was also used for GWAS on seed protein and oil content. All the associated SNPs detected by ECMLM overlapped with that from CMLM indicating that the result of the association mapping in the present study was reliable (Table S1).

A total of 199 soybean genes were found in the 50 kbp flanking region of each peak SNP (Table S2). Of these genes, 38 genes had no functional annotation. The other 161 genes were classified into 21 groups and might participate in 21 kinds of biological processes (Fig. 5). Genes involved in major CHO metabolism, lipid metabolism, N-metabolism, and amino acid metabolism might affect soybean oil and/or protein content (Table 4).

4. Discussion

Soybean seed oil and protein constitute complex quantitative traits that exhibit significant environmental influence and are governed by multiple genetic loci, each mostly displaying minor effects [36]. Loci exhibiting a minor effect and poor repeatability are often difficult to locate. To date, numerous soybean seed protein and oil QTLs have been effectively tagged using a variety of molecular marker systems, including simple sequences repeats (SSRs), random amplified polymorphic DNA (RAPD), restriction fragment length polymorphisms (RFLPs), and SNPs (http://www.soybase.org) based on linkage analysis, mainly using cross populations. Using a diverse collection of soybean germplasm samples, the QTLs associated with seed protein and oil content in this study were identified using a GWAS mapping approach and sequence-based SNP maps. Thirty-one QTLs were identified in total. Of these, SNP markers rs53140888 on Chr 02, rs19485676 on Chr 13, and rs24787338 on Chr 20, which were discovered as essential to protein content, were detected in both 2015 and 2016, indicating that the above three SNPs exhibited genetic stability. The SNP markers rs15774585 on Chr 07, identified as fundamental to oil content, were also stable in 2015 and 2016. Of these four stable SNP markers, rs19485676 overlapped with the reported QTL 'Prot26-13' associated with soybean seed protein content [37], and rs24787338 overlapped with six previously reported QTLs, including three seed protein content-related QTLs ('Prot1-1', 'Prot1-2', and 'Prot34-11') [38, 39] and three seed oil content-related QTLs ('Oil2-1', 'Oil2-2', and 'Oil24-30') [34, 40]. The other two stable SNP markers, rs53140888 and rs15774585, constituted novel loci underlying soybean seed protein content and were identified for the first time in the present study.

We discovered a total of 23 SNPs that were correlated with either soybean seed protein or oil content in only 2015 or 2016. Of them, nine were found to overlap with reported QTLs underlying soybean seed protein or oil content, suggesting that these SNPs showed reproducibility in different independent experiments. For seed oil content, the SNP marker rs35315668 overlapped with the oil content-related QTL 'Oil4-1' [41], and the SNP markers rs12328685 and rs49097495 overlapped with the protein content-related QTLs 'Prot34-2' and 'Prot16-2', respectively [1, 42]. For seed protein content, three SNP markers (rs4187467, rs6121092, and rs18150116) overlapped with protein content-related QTLs ('Prot12-2' and 'Prot26-13'), respectively [14, 43]. Another two SNP markers, rs29457452 and rs29979450, overlapped with both protein and oil content-related QTLs. Oil and protein content were found to be negatively correlated [2, 44]. Previous studies documented a series of pleiotropic QTLs underlying soybean seed and oil compositions [42]. In the present study, rs12328685, rs49097495, rs29457452, and rs29979450 were pleiotropic for soybean seed protein and oil content. It is therefore essential to clarify the effects of the loci for the two target traits before initiating MAS programs.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.01.004.

Acknowledgements

This study was conducted in the Key Laboratory of Soybean Biology of the Chinese Education Ministry, Agriculture Research System of China, and the key Laboratory of Northeastern Soybean Biology and Breeding/Genetics of Chinese Agriculture Ministry, and was financially supported by the National Key R&D Project for Crop Breeding (2017YFD0101302, 2016YFD0100304), Chinese National Natural Science Foundation (31571693), Chinese National Natural Science Foundation (31501251) and the Postdoctoral Fund in Heilongjiang Province (LBH-Z15017).

References

[1] Li Y-h, J.C. Reif, Hong H-l, Li H-h, Liu Z-x, Y.-s. Ma, J. Li, Y. Tian, Y.-f. Li, W.-b. Li, et al., Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions, Plant Sci. 266 (Supplement C) (2018) 95–101.

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

[3] J.W. Burton, C.A. Brim, Recurrent selection in soybeans. III. Selection for increased percent oil in seeds, Crop Sci. 21 (1) (1981).

[4] V. Channamallikarjuna, H. Sonah, M. Prasad, G.J.N. Rao, S. Chand, H.C. Upreti, N.K. Singh, T.R. Sharma, Identification of major quantitative trait loci qSBR11-1 for sheath blight resistance in rice, Mol. Breed. 25 (1) (2010) 155–166.

[5] R. Deshmukh, H. Sonah, G. Patil, W. Chen, S. Prince, R. Mutava, T. Vuong, B. Valliyodan, H.T. Nguyen, Integrating omic approaches for abiotic stress tolerance in soybean, Front. Plant Sci. 5 (2014) 244.

[6] L.P. Manavalan, S.J. Prince, T.A. Musket, J. Chaky, R. Deshmukh, T.D. Vuong, L. Song, P.B. Cregan, J.C. Nelson, J.G. Shannon, Identification of novel QTL governing root architectural traits in an interspecific soybean population, PLoS One 10 (3) (2015) e0120490.

[7] S. Salvi, R. Tuberosa, To clone or not to clone plant QTLs: present and future challenges, Trends Plant Sci. 10 (6) (2005) 297.

[8] Y. Cao, S. Li, Z. Wang, F. Chang, J. Kong, J. Gai, T. Zhao, Identification of major quantitative trait loci for seed oil content in soybeans by combining linkage and genome-wide association mapping, Front. Plant Sci. 8 (2017).

[9] C. Zhu, M. Gore, E.S. Buckler, J. Yu, Status and prospects of association mapping in plants, Plant Genome J. 1 (1) (2008) 5.

[10] H. Sonah, L. O'Donoughue, E. Cober, I. Rajcan, F. Belzile, Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean, Plant Biotechnol. J. 13 (2) (2015) 211.

[11] N.A. Baird, P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko, E.A. Johnson, S.N.P. Rapid, Discovery and genetic mapping using sequenced RAD markers, PLoS One 3 (10) (2008) e3376.

[12] S. Wang, E. Meyer, J.K. Mckay, M.V. Matz, 2b-RAD: a simple and flexible method for genome-wide genotyping, Nat. Methods 9 (8) (2012) 808.

[13] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A. Robust, Simple genotyping-by-sequencing (GBS) approach for high diversity species, PLoS One 6 (5) (2011) e19379.

[14] X. Sun, D. Liu, X. Zhang, W. Li, H. Liu, W. Hong, C. Jiang, N. Guan, C. Ma, H. Zeng, et al., SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing, PLoS One 8 (3) (2013) e58700.

[15] S. Atwell, Y.S. Huang, B.J. Vilhjalmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A.M. Tarone, Hu TT, et al., Genome-wide association study of 107 phenotypes in Arabidopsis Thaliana inbred lines, Nature 465 (7298) (2010) 627–631.

[16] X. Huang, X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, Genome-wide association studies of 14 agronomic traits in rice landraces, Nat. Genet. 42 (11) (2010) 961.

[17] H. Li, Z. Peng, X. Yang, W. Wang, J. Fu, J. Wang, Y. Han, Y. Chai, T. Guo, N. Yang, Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels, Nat. Genet. 45 (1) (2013) 43–50.

[18] S. Stracke, G. Haseneyer, J.B. Veyrieras, H.H. Geiger, S. Sauer, A. Graner, H.P. Piepho, Association Mapping Reveals Gene Action and Interactions in the Determination of Xowering Time in Barley, (2009).

[19] C. Sauvage, V. Segura, G. Bauchet, R. Stevens, P.T. Do, Z. Nikoloski, A.R. Fernie, M. Causse, Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits, Plant Physiol. 165 (3) (2014) 1120.

[20] M.A. Newell, D. Cook, N.A. Tinker, J.L. Jannink, Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies, Theor. Appl. Genet. 122 (3) (2011) 623.

[21] G.P. Morris, P. Ramu, S.P. Deshpande, C.T. Hash, T. Shah, H.D. Upadhyaya, O. Rieralizarazu, P.J. Brown, C.B. Acharya, S.E. Mitchell, Population genomic and genome-wide association studies of agroclimatic traits in sorghum, Proc. Natl. Acad. Sci. U. S. A. 110 (2) (2013) 453.

[22] S. Mamidi, R.K. Lee, J.R. Goos, P.E. Mcclean, Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (glycine max), PLoS One 9 (9) (2014) e107469.

[23] J. Sun, N. Guo, J. Lei, L. Li, G. Hu, H. Xing, Association mapping for partial resistance to Phytophthora sojae in soybean (Glycine max (L.) Merr.), J. Genet. 93 (2) (2014) 355.

[24] I. Elmer, S. Humira, B. François, Association mapping of QTLs for sclerotinia stem rot resistance in a collection of soybean plant introductions using a genotyping by sequencing (GBS) approach, BMC Plant Biol. 15 (1) (2015) 5.

[25] X. Zhao, Y. Han, Y. Li, D. Liu, M. Sun, Y. Zhao, C. Lv, D. Li, Z. Yang, L. Huang, Loci and candidate gene identification for resistance to Sclerotinia sclerotiorum in soybean (Glycine max L. Merr.) via association and linkage maps, Plant J. 82 (2) (2015) 245–255.

[26] Y. Bao, T. Vuong, C. Meinhardt, P. Tiffin, R. Denny, S. Chen, H.T. Nguyen, J.H. Orf, N.D. Young, Potential of association mapping and genomic selection to explore PI 88788 derived soybean cyst nematode resistance, Plant Genome 7 (3) (2014) 2840–2854.

[27] Y. Han, X. Zhao, G. Cao, Y. Wang, Y. Li, D. Liu, W. Teng, Z. Zhang, D. Li, L. Qiu, Genetic characteristics of soybean resistance to HG type 0 and HG type 1.2.3.5.7 of the cyst nematode analyzed by genome-wide association mapping, BMC Genomics 16 (1) (2015) 1–11.

[28] T.D. Vuong, H. Sonah, C.G. Meinhardt, R. Deshmukh, S. Kadam, R.L. Nelson, J.G. Shannon, H.T. Nguyen, Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean, BMC Genomics 16 (1) (2015) 1–13.

[29] Z. Wen, R. Tan, J. Yuan, C. Bales, W. Du, S. Zhang, M.I. Chilvers, C. Schmidt, Q. Song, P.B. Cregan, Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean, BMC Genomics 15 (1) (2014) 809.

[30] E.Y. Hwang, Q. Song, G. Jia, J.E. Specht, D.L. Hyten, J. Costa, P.B. Cregan, A genome-wide association study of seed protein and oil content in soybean, BMC Genomics 15 (1) (2014) 1.

[31] X. Wu, C. Ren, T. Joshi, T. Vuong, D. Xu, H. Nguyen, SNP discovery by high-throughput sequencing in soybean, BMC Genomics 11 (1) (2010) 469.

[32] X. Sun, D. Liu, X. Zhang, W. Li, H. Liu, W. Hong, C. Jiang, N. Guan, C. Ma, H. Zheng, SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing, PLoS One 8 (3) (2013) e58700.

[33] Y. Han, X. Zhao, D. Liu, Y. Li, D.A. Lightfoot, Z. Yang, L. Zhao, G. Zhou, Z. Wang, L. Huang, Domestication footprints anchor genomic regions of agronomic importance in soybeans, The New Phytol. 209 (2) (2016) 871–884.

[34] J. Schmutz, S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, et al., Genome sequence of the palaeopolyploid soybean, Nature 463 (7278) (2010) 178–183.

[35] A.E. Lipka, F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, M.A. Gore, E.S. Buckler, Z. Zhang, GAPIT: genome association and prediction integrated tool, Bioinformatics 28 (18) (2012) 2397–2399.

[36] M. Akond, S. Liu, M. Boney, S.K. Kantartzi, K. Meksem, N. Bellaloui, D.A. Lightfoot, M.A. Kassem, Identification of quantitative trait loci (QTL) underlying protein, oil, and five major fatty acids contents in soybean, Am. J. Plant Sci. 5 (1) (2014) 158–167.

[37] Y. Reinprecht, V.W. Poysa, K. Yu, I. Rajcan, G.R. Ablett, K.P. Pauls, Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (Glycine max (L.) Merrill) germplasm, Genome 49 (12) (2006) 1510–1527.

[38] B.W. Diers, P. Keim, W.R. Fehr, R.C. Shoemaker, RFLP analysis of soybean seed protein and oil content, Theor. Appl. Genet. 83 (5) (1992) 608–612.

[39] W. Lu, Z. Wen, H. Li, D. Yuan, J. Li, H. Zhang, Z. Huang, S. Cui, W. Du, Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean, Theor. Appl. Genet. 126 (2) (2013) 425.

[40] Z. Qi, Q. Wu, X. Han, Y. Sun, X. Du, C. Liu, H. Jiang, G. Hu, Q. Chen, Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes, Euphytica 179 (3) (2011) 499–514.

[41] E.C. Brummer, G.L. Graef, J. Orf, J.R. Wilcox, R.C. Shoemaker, Q.T.L. Mapping, For seed protein and oil content in eight soybean populations, Crop Sci. 37 (2) (1997) 370–378.

[42] A. Chapman, V.R. Pantalone, A. Ustun, F.L. Allen, D. Landauellis, R.N. Trigiano, P.M. Gresshoff, Quantitative trait loci for agronomic and seed quality traits in an F2 and F4:6 soybean population, Euphytica 129 (3) (2003) 387–393.

[43] J.E. Specht, K. Chase, M. Macrander, G.L. Graef, J. Chung, J.P. Markwell, M. Germann, J.H. Orf, K.G. Lark, Soybean response to water: a QTL analysis of drought tolerance, Crop Sci. 41 (2) (2001) 493–509.

[44] J. Wilcox, Interrelationships among seed quality attributes in soybean, Crop Sci. 41 (1) (2001) 11–14.