Theor Appl Genet (2015) 128:1061–1072
DOI 10.1007/s00122-015-2490-4

ORIGINAL PAPER

Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]

Ying Hu Zhang · Mei Feng Liu · Jian Bo He · Yu Feng Wang · Guang Nan Xing · Yan Li · Shou Ping Yang · Tuan Jie Zhao · Jun Yi Gai

Received: 15 November 2014 / Accepted: 24 February 2015 / Published online: 10 March 2015

Abstract

Key message **After two cycles of marker-assisted breeding on three loci, lines with transgressive segregation of 8.22–9.32 % protein content were developed based on four original soybean parents with 35.35–44.83 % protein content.**

Abstract Marker-assisted breeding has been an innovative approach in conventional breeding, which is to be further demonstrated, especially for quantitative traits. A study on continuous transgressive breeding for seed protein content (SPC) in soybean using marker-assisted procedures is reported here. The SPC of the recombinant inbred line (RIL) population XG varied in 38.04–47.54 % under five environments with P_1 of 35.35 %, P_2 of 44.34 % and total heritability of 89.11 %. A transgressive segregant XG30 with SPC 45.53 % was selected for further improvement. The linkage mapping of XG showed its genetic constitution composed of five additive QTL (32.16 % of phenotypic variation or PV) and two pairs of epistatic QTL (2.96 % PV) using 400 SSR markers with the remnant heritability 53.99 % attributed to the undetected collective of minor QTL. Another transgressive segregant WT133 with SPC 48.39 % was selected from the RIL population WT (44.83 % SPC for both parents). XG30 and WT133 were genotyped on the three major additive QTL (*Prot-08-1*, *Prot-14-1* and *Prot-19-2*) as $A_2A_2B_2B_2L_1L_1$ and $A_1A_1B_1B_1L_2L_2$, respectively. From WT133×XG30, surprising transgressive progenies were obtained, among which the recombinants with all three positive alleles $A_2\_B_2\_L_2\_$ performed the highest SPC, especially that of *Prot-08-1*. The five F_2-derived superior families showed their means higher than the high parent value in $F_{2:3}$ and $F_{2:4}$ and more transgressive effect in $F_{2:5:6}$, with the highest as high as 54.15 %, or 4.82 and 9.32 % more than WT133 and its original high parent, respectively. This study demonstrated the efficiency of marker-assisted procedure in breeding for transgressive segregation of quantitative trait.

Communicated by I. Rajcan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2490-4) contains supplementary material, which is available to authorized users.

Y. H. Zhang · M. F. Liu · J. B. He · Y. F. Wang · G. N. Xing · Y. Li · S. P. Yang · T. J. Zhao (✉) · J. Y. Gai (✉)
Soybean Research Institute, Nanjing Agricultural University; National Center for Soybean Improvement, Ministry of Agriculture; Key Laboratory of Biology and Genetic Improvement of Soybean, Ministry of Agriculture; National Key Laboratory for Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, Jiangsu, China
e-mail: tjzhao@njau.edu.cn

J. Y. Gai
e-mail: sri@njau.edu.cn

Introduction

Soybean is the most important vegetative protein and oil source in the world due to its about 40 % protein and 20 % oil content. It has been reported that soybean accounts for 68 % of the world protein consumption (http://www.soystats.com). To increase the seed protein content (SPC) is one of the major breeding targets for soybean.

In Chinese soybean landraces and commercial cultivars, SPC ranged from 35 to 50 % according to Liu et al. (2009) and Hwang et al. (2014). The soybean SPC is known as a quantitatively inherited trait. To identify the SPC QTL/gene, linkage mapping has been widely used based on a number of mapping procedures. Su et al. (2010) used a simulation to study the fitness of various mapping procedures on four different genetic models for recombinant inbred lines (RIL). They

suggested a mapping strategy for RIL data with unknown genetic model, i.e., a full model procedure scanning, such as QTLNetwork v2.0, followed by verification with other procedures corresponding to the results from full model scanning.

A great number of SPC QTL have been reported over the past two decades, among them 136 QTL were listed in the USDA Soybean Genome Database (SoyBase, http://www.soybase.org). Several of these QTL have been identified for four or more times at identical or similar positions on chromosomes in different populations, such as Gm04: 43.6–47.7 Mb, Gm05: 39.7–41.4 Mb, Gm07: 4.2–9.6 Mb, Gm08: 5.8–10.2 Mb, Gm14: 4.8–9.6 Mb, Gm15: 0.0–7.5 Mb, Gm18: 47.9–54.0 Mb, Gm19: 35.5–42.1 Mb and Gm20: 2.1–34.2 Mb (Supplementary Table S1).

Marker-assisted breeding based on QTL/gene mapping has been considered as an innovative approach for precision plant breeding in the twenty-first century (Collard and Mackill 2008). Sebolt et al. (2000) used an allele from wild soybean to increase seed protein content in soybean. Sebastian et al. (2010) reported a marker-assisted breeding approach that involved the sub-lining of existing soybean elite cultivars derived from single F_3 or F_4 plants, and succeeded to increase accuracy of preliminary yield test.

"Breeding by design" as a concept based on QTL/gene mapping was firstly described by Peleman and van der Voort (2003). Gai et al. (2012) indicated that the "QTL-allele matrices" of the breeding traits could be established through QTL/gene association mapping integrated with linkage mapping, based on the results, optimal cross design and progeny selection could be performed. However, though the marker-assisted breeding has been extensively highlighted, it is still to be further demonstrated, especially for quantitatively inherited traits.

The present study is to report the results on continuously transgressive breeding for SPC using marker-assisted procedures. Two transgressive segregants were selected from two respective RIL populations, and were genotyped on their major additive QTL, respectively. From the cross between the two transgressive families, further transgressive progenies were obtained through the marker-assisted procedure. This result was used to demonstrate the potential of transgressive segregation and the usefulness of marker-assisted breeding procedure in cross design and progeny selection for quantitative traits.

Materials and methods

First cycle recombination experiments

Plant materials

The NJRSXG (XG) containing 147 lines is a F_6-derived RIL population from a cross between low-protein line Xianjin 2 (35.35 %) and high-protein line Gantai-2-2 (44.34 %) using single-seed descent from F_2 to F_5, where Xianjin 2 of Maturity Group IV (MG IV) is a selection from an unknown landrace in Anhui province, Gantai-2-2 (MG III) is a breeding line derived from a cross between the landraces Ganjiangnan and Taixingheidou in Jiangsu province.

Another RIL population is NJRSWT (WT) containing 142 F_6-derived lines from a cross between a cultivar Wan 82-178 (MG IV, 44.83 % SPC) and a landrace TSBPHDJ (MG V, 44.83 % SPC) using single-seed descent from F_2 to F_5.

Field experiments and seed protein content measurement

NJRSXG and its parents were tested in a randomized complete block (RCB) hill plot experiments with 2 replications at Jiangpu Experimental Station of Nanjing Agricultural University, Nanjing, China in 2009, 2010, 2010 (planted 1 month later than the former) and 2011; and 2 replications at Lishui Experimental Station, Nanjing, China in 2009. The hill plots were planted with 0.7 × 0.8 m spacing, each plot thinned to a final stand of six seedlings per plot. The planting dates were June 19 in 2009JP, June 26 in 2009LS, June 23 in 2010JP, July 23 in 2010JPL and June 27 in 2011JP, where 2009JP, 2010JP, 2010JPL, 2011JP and 2009LS are codes of the environments, respectively, for example, 2009 is the year, JP is the location of Jiangpu, LS is the location of Lishui and JPL is the location of Jiangpu but at a late planting date. The same experiments were for NJRSWT, except only tested under two environments, 2009JP and 2009LS.

A specimen of 20 g seeds for each replication, each line in NJRSXG and NJRSWT were milled with a 1095 Knifetec sample mill (FOSS Tecator, Denmark), then NIR spectroscopy analysis was performed with VECTOR22/N (BRUKER, German), and at last the seed protein content were converted with the calibration model developed by Wang (2011).

Genotyping and construction of genetic linkage map

The construction of genetic linkage map for NJRSXG has been previously described by Wang (2009). The NJRSXG map consists of 400 SSR markers on 23 linkage groups (LG) with three chromosomes split into two LGs, covering 1412.9 cM with an average of 3.9 cM between markers. Most of the SSR markers of NJRSXG were corresponding to the physical map (Schmutz et al. 2010) based on the marker primers.

QTL mapping procedure

Following the mapping strategy of Su et al. (2010), the mixed-model-based composite interval mapping (MCIM)

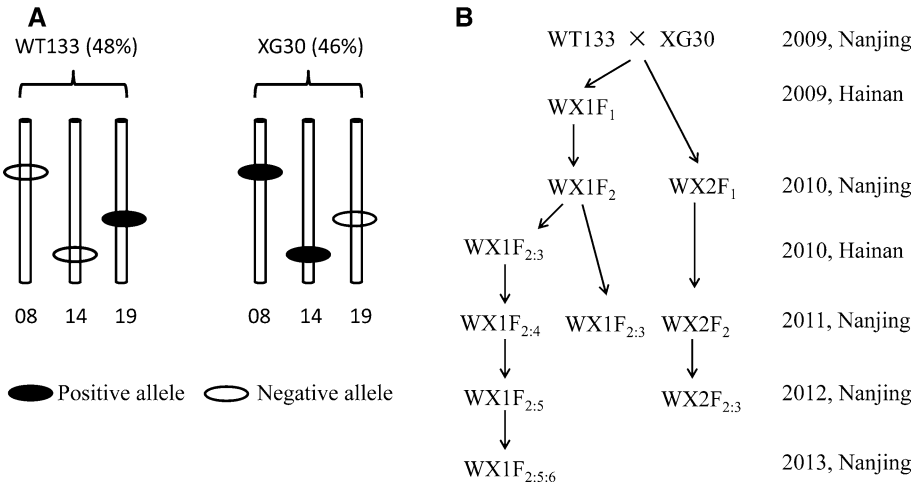

Fig. 1 The diagram of genetic structure of protein QTL in WT133 ($A_1A_1B_1B_1L_2L_2$) and XG30 ($A_2A_2B_2B_2L_1L_1$) (**a**) and the diagram of the developing process of the WX progenies (**b**)

of QTLNetwork V2.1 (Yang et al. 2008) was used to detect additive QTL, additive × additive epistatic QTL pairs, additive QTL × year, and epistatic QTL pair × year interaction. Significant QTL were obtained from *F*-statistic profile, the critical *F* value was calculated by permutation tests of 1000 times at genome-wise significance level of 0.05.

Statistical analysis and partition of genetic variation of the population

A joint analysis of variance (ANOVA) was conducted for NJRSXG and NJRSWT, respectively, using PROC GLM of SAS 9.1 (SAS Institute, Cary, NC) with Line, Environment, Replication and Line × Environment as random effects. The heritability (h^2) of SPC was calculated as $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2/r)$ for individual environments and $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_{gy}^2/n + \sigma_e^2/\mathrm{nr})$ for combined data, where σ_g^2, σ_{gy}^2 and σ_e^2 are Genotype, Genotype × Environment, and Error components of variance estimated from the expected mean squares in ANOVA, respectively, n is the number of environments, and r is the number of replications (Hanson et al. 1956). In calculation of expected mean squares, n and r were estimated with SAS procedures due to missing data and the replication numbers were different among environments. Genetic coefficient of variation (GCV) was calculated as GCV $= \sigma_g/\mu$, where μ was the population mean.

The genetic portion of the phenotypic variation (PV) from ANOVA was larger than those of the QTL detected in mapping procedures, as Xing et al. (2012) indicated, the difference of $h^2 - h_{(a)}^2 - h_{(aa)}^2$, was considered the part due to undetected QTL (missing heritability), and designated as collective unmapped QTL, where h^2 is the heritability estimated from ANOVA, and $h_{(a)}^2$ and $h_{(aa)}^2$ are additive and epistatic contributions, respectively, obtained from the QTLNetwork V2.1 mapping procedure.

Second cycle recombination experiment

Plant materials

A transgressive segregant XG30 with SPC 45.53 % was selected from XG population and so was WT133 with SPC 48.39 % from WT population for further improvement. WT133 and XG30 were crossed to develop WX population, including WX1 at first and then WX2 sub-populations as well as their further derived families as shown in Fig. 1a while the progeny development procedure is shown in Fig. 1b. Here WX2 was designed as a replicated cross for verification of the WX1 results. Based on the mapping results, XG30 and WT133 were genotyped on the three major additive QTL as $A_2A_2B_2B_2L_1L_1$ and $A_1A_1B_1B_1L_2L_2$, respectively (corresponding to *Prot-08-1*, *Prot-14-1* and *Prot-19-2*, see the "Results" section).

Field experiments and seed protein content measurement

The F_2, $F_{2:3}$, $F_{2:4}$, $F_{2:5}$ and $F_{2:5:6}$ of the cross WX were tested at Jiangpu Experimental Station of Nanjing Agricultural University, Nanjing, China. WX1 $F_{2:3}$, WX1 $F_{2:4}$, WX2 $F_{2:3}$ and WX1 $F_{2:5:6}$ were planted in a single line plot with 4 m long and 0.5 m apart in a RCB design with two replications.

The SPC of families in WX1 $F_{2:3}$, WX1 $F_{2:4}$, WX2 $F_{2:3}$, WX1 $F_{2:5:6}$ along with their parents were measured with a near infrared transmittance spectroscopy (FOSS Infratec 1241 Grain Analyzer, Foss Tecator, Sweden). The instrument used in the second cycle recombination experiment was different from that one used in the first cycle experiment. There was a systematic bias (0.6 %) between the two instruments. Therefore, the data of this cycle experiment were corrected to be comparable with those of the first cycle experiment.

Genotyping and data analysis

Genomic DNA was extracted from the sample leaves of soybean seedlings using the CTAB method (Murray and Thompson 1980). The genotype of F_2 plants of WX and the selected $F_{2:5}$ in WX were determined with SSR and presence/absence variation (PAV) markers flanking the detected QTL in XG through polyacrylamide gel electrophoresis. The primers of the SSR and PAV were developed according to Song et al. (2010) and our unpublished data, respectively.

For multiple test among the genotypes in $F_{2:3}$ and $F_{2:4}$ of WX1 and WX2, Duncan's new multiple range test was used.

For weighting the genetic contribution of each of the three major loci in WX, a linear model was fitted to assess the association between a marker genotype and seed protein content using PROC GLM of SAS 9.1, where SPC was used as the response variable and only the polymorphic homozygote was used as independent variable for each of the loci (*Prot-08-1*, *Prot-14-1* and *Prot-19-2*), respectively. Based on it, each locus contribution was estimated from the r^2 value in the regression analysis.

Marker-assisted breeding procedure

The major markers of SPC were detected mainly at the first cycle recombination. Based on the SPC marker information, XG30 and WT133, the parents of the second cycle recombination were selected from XG and WT, respectively. Then the progenies of WT133×XG30 (WX) were selected according to their marker information integrated with their phenotypic performance. Therefore, marker-assisted breeding in the present study was carried out in two stages, the marker-assisted parent selection and marker-assisted progeny selection.

Results

First cycle transgressive segregation of seed protein content in NJRSXG and NJRSWT and their genetic constitution

Genetic variability of seed protein content in NJRSXG and NJRSWT

Table 1 showed the frequency distribution of SPC in the RIL population NJRSXG. It ranged from 38.04 to 47.54 % based on the entry mean over five environments and from 35.89 to 49.10 % based on individual environments. A large transgressive segregation after genetic recombination came out, especially at the higher SPC side. In ANOVA, significant variations among lines and lines × environments were

Table 1 Frequency distribution and descriptive statistics of seed protein content in NJRSXG (%)

Env.	P_1	P_2	Class mid-point value												Mean	Min.	Max.	*GCV* (%)	h^2 (%)
			<37.0	37.5	38.5	39.5	40.5	41.5	42.5	43.5	44.5	45.5	46.5	≥47.0	$\sum f$				
09JP	36.23	45.96	0	0	0	3	17	25	30	28	19	17	4	1	144	43.02	39.16	47.77	3.39	79.01
09LS	35.73	43.54	1	0	3	12	23	29	38	15	17	4	1	2	145	42.15	36.62	47.52	3.84	82.97
10JP	36.26	43.86	1	3	4	12	24	27	22	21	16	9	2	5	146	42.29	36.61	49.10	4.66	81.55
10JPL	32.93	43.12	4	9	19	18	30	23	17	11	8	4	1	2	146	40.89	35.89	47.39	5.16	87.49
11JP	35.59	45.24	0	1	1	2	7	22	29	32	30	12	7	3	146	43.30	37.36	48.44	3.83	86.46
Mean	35.35	44.34	0	0	2	5	23	35	28	32	11	7	2	1	146	42.33	38.04	47.54	3.59	89.11

The same is for the latter tables

P_1 Xianjin 2, P_2 Gantai 2-2, *Env.* environment, *Min.* Minimum, *Max.* Maximum, *GCV* genotypic coefficient of variation, h^2 heritability value calculated from ANOVA

09JP experiment in JiangPu Station in 2009, *09LS* experiment in LiShui Station in 2009, *10JP* experiment in JiangPu Station in 2010, *10JPL* experiment in JiangPu Station in 2010 but planted 1 month later than the normal summer planting date, *11JP* experiment in JiangPu Station in 2011

Table 2 Frequency distribution and descriptive statistics of seed protein content in NJRSWT (%)

Env.	P_1	P_2	Class mid-point value												Mean	Min.	Max.	*GCV* (%)	h^2 (%)
			39.5	40.5	41.5	42.5	43.5	44.5	45.5	46.5	47.5	48.5	49.5	50.5	$\sum f$					
09JP	45.54	44.10	2	8	13	22	32	20	12	13	11	8	1	0	142	44.19	39.11	49.44	4.66	86.87
09LS	44.11	45.56	0	0	6	6	17	29	37	27	12	5	2	1	142	45.25	41.42	50.32	3.45	84.78
Mean	44.83	44.83	0	1	7	15	26	37	28	13	8	5	2	0	142	44.72	40.67	49.53	3.62	87.09

P_1 Wan 82-178, P_2 TSBPHDJ

Table 3 QTL conferring protein content detected in NJRSXG population

QTL	Flanking markers	Chr.	LG	Site		Additive QTL		Epistatic QTL [aa, h^2aa (%)]	References
				cM	Mb	a	h_a^2 (%)		
Prot-01-1	Satt532-Satt179	Gm01	D1a	12.6	28.2–38.8			*Prot-15-1* (−0.30, 1.53)	
Prot-02-1	Satt579-GNE540	Gm02	D1b	14.2	17.4–19.4			*Prot-19-1* (−0.27, 1.43)	Chen et al. (2007)
Prot -08 -1	**Sat_199-Satt525**	**Gm08**	*A2*	**2.2**	**15.1–17.1**	**0.78**	**8.90**		
Prot-08-2	Satt437-Satt329	Gm08	A2	24.4	18.9–21.2	−0.59	2.45		
Prot -14 -1	**Satt556-Satt020**	**Gm14**	**B2**	**10.3**	**39.6–42.0**	**−0.57**	**11.06**		Liang et al. (2010)
Prot-15-1	GNB223-GNB186	Gm15	E	15.1	41.1–42.3			*Prot-01-1* (−0.30, 1.53)	Chen et al. (2007)
Prot-16-1	Satt596-Sat_151	Gm16	J	51.3	14.0–18.4	0.29	1.47		
Prot-19-1	Sat388-GNE047	Gm19	L	24.2	31.2–36.8			*Prot-02-1* (−0.27, 1.43)	Diers et al. (1992), Tajuddin et al. (2003)
Prot - 19 -2	**GNE047-Satt481**	**Gm19**	**L**	**38.1**	**36.8–40.1**	**0.53**	**8.28**		Tajuddin et al. (2003), Orf et al. (1999)
Sum							32.16	2.96	

The bold QTL are major ones used for further study

In Site column, the cM number is the position with highest *F* on NJRSXG genetic linkage map, while the Mb numbers are the physical position (genome assembly 1 annotation version 1.1) of the closest flanking markerson the public soybean physical map (Schmutz et al. 2010)

QTL the QTL detected with MCIM of QTLNetwork v2.1 is designated as *Prot-08-1*, where *Prot* means protein, *-08* represents Chromosome 8 and *-1* represents its order on the chromosome according to its physical position, *Chr.* chromosome, *LG* linkage group

observed with the *F* value of the former much larger than that of the latter (ANOVA tables omitted). The heritability estimated from the expected mean squares in ANOVA of the combined data was 89.11 % and those for the individual environments ranged in 79.01–87.49 %, which suggests that the genetic variation accounts for a major part of the phenotypic variance (PV) and the total contribution of the detected QTL can be up to 80–90 % of the PV.

In another RIL population NJRSWT, the SPC ranged in 40.67–49.53 %, with transgressive segregation as large as more than 4 % SPC at both sides in comparison with the SPC 44.83 % for both parents. The heritability of the whole experiment was 87.09 % while those of the individual environments were 84.78–86.87 % (Table 2).

QTL of seed protein content detected in NJRSXG

Table 3 shows the QTL mapping results for SPC in NJRSXG. Altogether, five additive QTL and two epistatic QTL pairs on seven chromosomes (linkage groups) were detected. Among the five additive QTL, three positive alleles came from Gantai2-2 (high SPC parent), and two positive alleles came from Xianjin 2 (low SPC parent). The QTL of *Prot-08-1*, *Prot-14-1* and *Prot-19-2* accounted for 8.90, 11.06 and 8.28 % of PV, respectively, therefore were considered as major additive QTL, while the QTL of *Prot-08-2* and *Prot-16-1* accounted for 2.45 and 1.47 % of PV, respectively, and were considered as small additive QTL.

Two epistatic QTL pairs (*Prot-01-1 and Prot-15-1; Prot-02-1 and Prot-19-1*) were detected with their contribution 1.53 and 1.43 %, respectively, relatively smaller than those of the additive QTL. All the four loci were located on different chromosomes with epistatic effects only but no additive effects.

Five of the total nine QTL were located on or near the region of previously reported region protein QTL (Table 3), indicating the protein QTL detected in different population could corresponding. While the other four, i.e. *Prot-01-1*, *Prot-08-1*, *Prot-08-2* and *Prot-16-1* were newly found in the present study.

Table 4 Components of genetic contribution to the total phenotypic variation of protein content in NJRSXG under multiple environments (%)

Genetic contribution				QTL × Env.				Error	Overall
Additive QTL	Epistatic QTL pair	UCM QTL	Total	Additive QTL × Env.	Epistatic QTL × Env.	UCM QTL × Env.	Total		
32.16 (5; 36.09)	2.96 (2; 3.32)	53.99 (60.59)	89.11	2.57	0.78	2.81	6.16	4.73	100

In parentheses of the additive QTL and epistatic QTL pair columns, the left number is the number of additive QTL or epistatic QTL pairs; the right number is the percentage of genetic variance explained by the respective QTL or QTL pairs, while the number outside of parentheses is their total contribution to the phenotypic variation. UCM QTL means undetected collective of minor QTL, the number in parenthesis is the percentage of genetic variance explained by UCM QTL and the number outside of parentheses is its contribution to the phenotypic variation (or missing heritability)

Total genetic contribution, QTL × Env. and Error were the heritability, genotype × environment and the error, respectively, obtained from ANOVA, while the additive QTL, epistatic QTL pair, additive QTL × Env. and Epistatic QTL × Env. contribution obtained from the mapping procedure QTLNetwork 2.1, respectively. The UCM QTL and the UCM QTL × Env. were obtained from the total genetic contribution subtracted with the genetic variation due to all detected additive and epistatic QTL and the QTL × Env. subtracted with additive QTL × Env. and Epistatic QTL × Env., respectively

Env. environment

Genetic dissection of seed protein content of the NJRSXG population

From ANOVA, the total phenotypic variance of the NJRSXG population was partitioned into its components of genetic contribution among lines, line by environment interactions, and random error. From QTL mapping, the genetic contribution of the additive QTL and that of epistatic QTL pairs were estimated. To integrate the two sets of results, the total phenotypic variation could be separated into different parts, including genetic part, QTL × environment interaction part and random environment or error part as shown in Table 4. Here, the total genetic variance proportion is equivalent to heritability value estimated from ANOVA. The additive QTL proportion and the epistatic QTL pair proportion are obtained from the mapping procedure MCIM of QTLNetwork and the difference between total genetic variance proportion and the identified QTL proportion most likely is due to a group of QTL undetected by the mapping procedure, which is designated as undetected collective of minor (UCM) QTL (Korir et al. 2011) or the missing heritability. In Table 4, the additive QTL × environment, and epistatic QTL × environment were obtained from the mapping procedure MCIM of QTLNetwork, and the difference between line × environment in ANOVA and the additive QTL × environment plus epistatic QTL × environment should be due to UCM QTL × environment interaction. The error variation is obtained from ANOVA directly. Therefore, the genetic contribution due to additive QTL, epistatic QTL pairs and undetected QTL collective, the QTL × environment contribution due to additive QTL × environment, epistatic QTL × environment and UCM QTL × environment, and the error variation should sum to 100 % of the PV.

In Table 4, 89.11 % of the PV was accounted for by genetic variation, in which five additive QTL explained 32.16 %, two epistaic QTL pair explained 2.96 %, and the remnant 53.99 % as the majority was explained by the UCM QTL. In addition, 6.16 % of the PV was accounted for by line × environment variation, in which additive QTL × environment, epistatic QTL × environment and the UCM QTL × environment explained 2.57, 0.78 and 2.81 % PV, respectively. The results showed the line × environment variation was relatively small, while among the genetic components, the UCM QTL were the most important, followed by the additive QTL and the epistatic QTL. That means in the QTL mapping procedure, there are many QTL with minor effects which could not be detected under the present population size, marker density and experimental precision.

Genetic structure of the transgressive segregants XG30 and WT133

To obtain further transgressive segregation on SPC, the best lines in both populations were selected and genotyped for the major additive QTL *Prot-08-1*, *Prot-14-1* and *Prot-19-2* according to NJRSXG map as shown in Table 5. Among the superior lines, WT133 with SPC 48.39 % was chosen from NJRSWT and XG30 with SPC 45.53 % from NJRSXG. WT133 has its genotypes on the three major QTL as $A_1A_1B_1B_1L_2L_2$. Here A, B and L represent *Prot-08-1, Prot-14-1* and *Prot-19-2* and the subscription 2 and 1 represent positive and negative allele on a locus, respectively, while XG30 has its genotypes on the three major QTL as $A_2A_2B_2B_2L_1L_1$. It is obvious that the two superior lines have their genotypes on the three major additive loci complementary to each other (Fig. 1a). It implies further transgressive segregation of SPC could be achieved in the cross between the two superior lines. Accordingly, the cross WT133 × XG 30 (or WX) was made and their progenies were developed as shown in Fig. 1b. By the way, the genotypes of the other

Table 5 The major QTL genotype of the high seed protein content lines in NJRSXG and NJRSWT

Population	Line	Seed protein content	Genotype				
			Prot-08-1	*Prot-14-1*	*Prot-19-2*	*Prot-08-2*	*Prot-16-1*
NJRSXG	Xianjin 2	35.35	A_1A_1	B_2B_2	L_1L_1	X_2X_2	Y_1Y_1
	Gantai 2-2	44.34	A_2A_2	B_1B_1	L_2L_2	X_1X_1	Y_2Y_2
	XG25	47.54	A_2A_2	B_2B_2	L_2L_2	X_2X_2	Y_1Y_1
	XG91	46.62	A_2A_2	B_2B_2	L_1L_1	X_2X_2	Y_1Y_1
	XG137	46.21	A_2A_2	B_2B_2	L_1L_1	X_2X_2	Y_2Y_2
	XG30	45.53	A_2A_2	B_2B_2	L_1L_1	X_2X_2	Y_1Y_1
NJRSWT	Wan 82-178	44.83	A_1A_1	B_1B_1	L_2L_2	–	–
	TSBPHDJ	44.83	A_1A_1	B_1B_1	L_2L_2	–	–
	WT77	49.53	A_1A_1	B_1B_1	L_2L_2	–	–
	WT125	49.07	A_1A_1	B_1B_1	L_2L_2	–	–
	WT117	48.76	A_1A_1	B_1B_1	L_2L_2	–	–
	WT133	48.39	A_1A_1	B_1B_1	L_2L_2	–	–

A, *B*, *L*, *X* and *Y* represent *Prot-08-1, Prot-14-1*, *Prot-19-2, Prot-08-2* and *Prot-16-1*, respectively, while the subscribe 1 means negative allele and subscribe 2 means positive allele. In NJRSXG, A_2, B_1, L_2, X_2 and Y_1 are from Gantai 2-2 while A_1, B_2, L_1, X_1 and Y_2 are from Xianjin 2. In NJRSWT, the genotypes of *A*, *B* and *L* were identified in a group of materials X and Y genotypes unknown, while the genotypes of the WT lines were inferred from the parents according to no allele difference on *A*, *B*, and *L* between the two parents

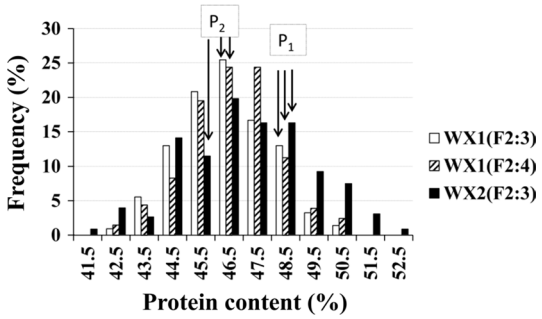

Fig. 2 The frequency distribution of seed protein content in WX population P_1 WT133, P_2 XG30

two loci (*Prot-08-2* and *Prot-16-1*) detected in XG of the superior lines with their parents were also listed in Table 5, but those for NJRSWT were unknown. However, it might leave some possibility for more recombination in WX since XG30 composed of one positive locus and one negative locus on *X* and *Y* loci if WT133 composed of the two loci in a way different from XG30.

Second cycle transgressive segregation of seed protein content in derived generations of WT133 × XG 30

Further transgressive segregation of seed protein content and comparisons among genotypes in F_2-derived generations of WX

Figure 2 shows the frequency distribution of SPC in WX, including WX1 ($F_{2:3}$), WX1 ($F_{2:4}$) and WX2 ($F_{2:3}$). The data confirmed that the parents WT133 (P_1) and XG30 (P_2) both had high SPC with the former higher than the latter. The range of SPC was 43.85–51.30 %, 43.65–51.60 % and 40.50–52.20 % in WX1$F_{2:3}$, WX1 $F_{2:4}$ and WX2 $F_{2:3}$, respectively. Obviously the distributions after the second cycle recombination have moved to the higher side with the highest segregations larger than those of the two RIL populations (or the first cycle of recombination). In other words, further transgressive segregation achieved through recombination of the two selected complementary superior lines.

The genetic effect of a QTL and its contribution to the population may vary among populations. WX is a new population composed of the genetic sources from four original parents. To evaluate the genetic effects of the three major loci and their genetic contribution in WX, their flanking SSR markers obtained in XG were used again to genotype the WX population. Those were Satt119, Satt377 and Satt525 for *Prot-08-1*, Satt020 for *Prot-14-1* and Satt481 for *Prot-19-2* which were not necessary the same but close to the markers (Table 3). In addition, nine presence/absence variation (PAV) markers near the three major loci (Table 6), recently reported by Wang et al. (2014), were added for exploring closer markers to obtain more precise evaluation of the allele effects and their genetic contribution.

To choose the best marker for each QTL from the 14 SSR and PAV, the simple regression analysis of SPC on each marker was conducted individually. The flanking marker with largest r^2 was chosen as the indicator of the QTL. It was done for WX1 and the selected markers were used for WX2. Then the lines in the respective F_2-derived populations were grouped according to their genotype for

Table 6 The flanking markers used for WX population

QTL	Marker	Chr.	Posi. (kb)	r^2	UP primer	Down primer
Prot-08-1	Z37	Gm08	13380	19.80	TCCATCGGGTCCATATATTA	TCCCTGTAAACAAACATAATGA
	Z50	Gm08	13905	35.91	TCGCAAACTCAATCTACAAA	AGTATTTTCTTTTCCTTTTATGT
	Z56	**Gm08**	**14193**	**43.83**	**GGTTTGTGTGGTGGTAAGAT**	**AACGTGATGGTTGGTTAGTT**
	Satt119	Gm08	15141	43.17	GCGCAATTTGACTATTTTTAGCTGTTG	GCGCGATATAAGATGATTTTTATTGAT
	Satt377	Gm08	16428	40.22	GCGATACACGTGACGAAGAGAATCATGC	GCGATAACATAATGGATTGACATAAAT
	Satt525	Gm08	17077	37.03	GCGCATAGCTTTCAGAGAAGTTT	CATTCACCAATCCTCATTAGA
Prot-14-1	Satt020	Gm14	42022	2.54	GAGAAAGAAATGTGTTAGTGTAA	CTTTTCCTTCTTATTGTTTGA
	Z69	Gm14	42051	2.76	GCTTAGATAGTTAAGATTTGCTT	CACATCAATGGATTATAAGGTG
	Z83	**Gm14**	**44895**	**2.98**	**TGGCAATTCTTATTCAGTCC**	**TGGCAATTTTTCGATTAGAC**
	Z98	Gm14	46855	1.62	CAAACAGCCAAACATAAACA	CTTCAATTTGAGTCCAGCTT
	Z100	Gm14	47102	1.60	TTCCCTTTCTTAAACCCTTT	ACAACGTGACACCTCATGTA
Prot-19-2	Satt481	Gm14	40127	2.96	GGGTTAACCGTCCACACATCTATT	GACGGTTTTAAACGGTAAGAAAAT
	Z135	Gm14	41040	3.05	ACCAAATTCATTACCACCAG	CAGTTTTTAGCGGAATCTTG
	Z136	**Gm19**	**41077**	**3.47**	**GTTTGTTTCAAAGTGGCATT**	**AGCGAATCTAGCGTACTAGC**

The position of the markers were based on genome assembly 1 annotation version 1.1

The r^2 was obtained from regression of SPC on a marker in WX1

The bold markers had the highest contribution to the phenotypic variation among the markers for each QTL, therefore, were used to represent the respective QTL for further analysis in WX population

Chr. Chromosome, *Posi.* Position

each QTL and the additive effect and dominance effect of each QTL were calculated according to $1/2(X_2X_2\text{-}X_1X_1)$ and $(X_1X_2\text{-}X_1X_1)\text{–}1/2(X_2X_2\text{-}X_1X_1)$, respectively, where *X* represents an allele of a QTL. The results showed that Z56, Z83 and Z136 were the best markers for *Prot-08-1*, *Prot-14-1* and *Prot-19-2*, respectively. Accordingly, the additive effect of the *Prot-08-1*, *Prot-14-1* and *Prot-19-2* were 1.11, 0.45 and 0.31 %, while the partial dominance effect of *Prot-08-1*, *Prot-14-1* and *Prot-19-2* were −0.31, 0.24 and −0.10 %, respectively, indicating the additive effect was larger than the dominance effect for each of the three major QTL or A_1, B_2 and L_1 were partially dominant over A_2, B_1 and L_2 in WX. Fig 3 showed the average difference between the two homozygous genotypes of *Prot-08-1*, *Prot-14-1* and *Prot-19-2* in WX-derived populations. The difference between the positive allele homozygotes and negative allele homozygotes on each of the three major QTL was significant at least in two F_2-derived populations, especially for *Prot-08-1* which was significant in all three F_2-derived populations. Among the three major QTL, the genetic contribution to the phenotypic variation, *Prot-08-1, Prot-14-1* and *Prot-19-2* explained 28.83–43.83 %, 2.98–7.45 % and 1.71–4.12 % of PV, respectively, among which *Prot-08-1* was the most important locus in WX population.

To demonstrate the best combination of the three loci for transgressive segregation, the lines of the F_2-derived populations were grouped according to their three-loci genotype. Table 7 showed the comparisons of SPC among different three-loci genotypes in the F_2-derived populations of WX. The trend was very obvious that the genotype with all positive alleles on the three loci performed the best on their SPC, those with two positive alleles on the three loci performed the next, while those with one positive allele and null positive allele on the three loci performed the least. Table 7 also showed that the genotypes of $A_2\_B_2\_L_2\_$, $A_2\_B_2\_L_1L_1$ and $A_2\_B_1B_1L_2\_$ in the first three rows had potential for positive transgressive segregation. This result demonstrates further the importance of the three major QTL for SPC. However, the data also showed that the genotypes with high SPC varied widely, some lines with less positive alleles on the three major QTL might perform better than those with more positive alleles and vice versa. It might be due to more interacted QTL involved with SPC and will be discussed in the Discussion section.

Further transgressive segregation of seed protein content in superior families derived from WT133 × XG30

From the above results, *Prot-08-1* was the most important QTL among the three major QTL in WX and the best families were derived mainly from F_2 plants with heterozygous A plus one of the heterozygous B or heterozygous L, from which further transgressive segregants could be obtained. Table 8 shows the segregation of five best families of WX1. All of their F_2 ancestor plants have heterozygous A. Their SPC averages at $F_{2:3}$ and $F_{2:4}$ were all higher than the high parent WT133 (49.93 and 49.74 % vs. 49.33 %). That means there must be segregants much higher than the high

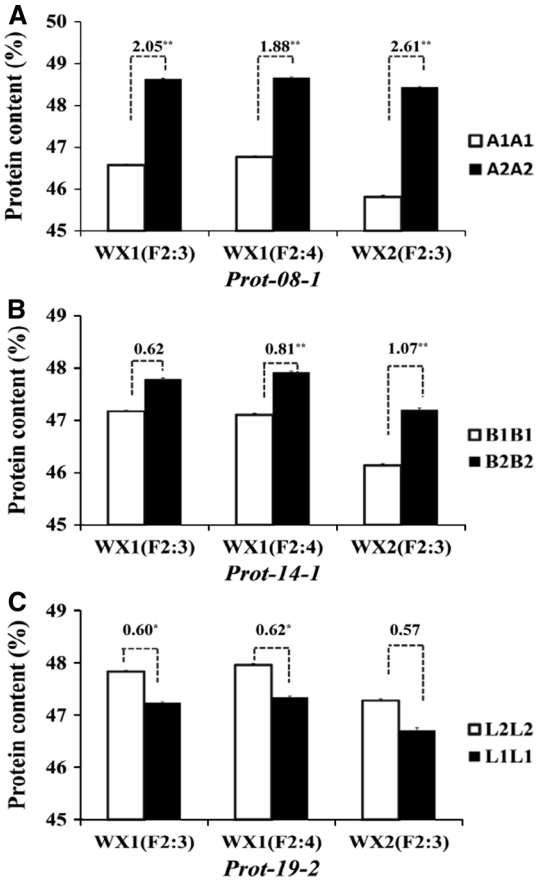

Fig. 3 The average difference of protein content between the two genotypes of the QTL of *Prot-08-1*(**a**), *Prot-14-1* (**b**) and *Prot-19-2* (**c**) in WX-derived populations to show the average allele effects of each locus. The number above the histogram is the average difference between the two genotypes on a locus. The phenotypic variation among the two homozygotes of *Prot-08-1* was 43.83, 30.19 and 28.83 % for WX1 ($F_{2:3}$), WX1 ($F_{2:4}$) and WX2 ($F_{2:3}$), respectively. The phenotypic variation among the two homozygotes of *Prot-14-1* was 2.98, 7.45 and 6.52 % for WX1 ($F_{2:3}$), WX1 ($F_{2:4}$) and WX2 ($F_{2:3}$), respectively. The phenotypic variation among the two homozygotes of *Prot-19-2* was 3.47, 4.12 and 1.71 % for WX1 ($F_{2:3}$), WX1 ($F_{2:4}$) and WX2 ($F_{2:3}$), respectively. *Asterisk* and *double asterisk* represent significant at $p < 0.05$ and $p < 0.01$, respectively

parent WT133 or must exist high transgressive segregants. Therefore, $F_{2:5:6}$ lines were derived and tested for their SPC.

Table 8 also showed the SPC distribution in $F_{2:5:6}$ of the five superior families. There was further segregation among the lines and the distributions move further to the high SPC side. The best lines derived from WX1-139 ($A_1A_2B_2B_2L_2L_2$) were WX1-139-95, WX1-139-68, WX1-139-99 and WX1-139-18 with genotype of all three loci positive effects ($A_2A_2B_2B_2L_2L_2$) having SPC 52.5, 52.1, 51.4 and 51.2 %, respectively. However, the maximum lines with SPC 53.05, 53.35, 54.15, 54.05 and 54.10 % (Table 8) derived from the five families were 3.72–4.82 % more than the high parent WT133 and 8.22–9.32 % more than their original high parent Wan 82-178 or TSBPHDJ.

Look at the $F_{2:5}$ genotypes of the superior families, all of them contain heterozygous A. That means the segregation of locus A should be of special importance. Fig 4 showed the differences of SPC between the two homozygotes A_2A_2 and A_1A_1 (*Prot-08-1*) on an average base for each of the five superior families. Among them, four showed significant difference (1.21–3.15 %) and only one was non-significant (0.69 %). The genetic contribution of *Prot-08-1* to PV ranged from 14.64 to 26.78 % for the four families while only 4.35 % for the exceptional one family, and at an average of 24.11 % for all the five superior families. It demonstrates the importance of *Prot-08-1* in determining the transgressive segregation of SPC and also implies the influence from the genetic background in the population.

In summarizing the results from the two cycles of transgressive segregation, at the first cycle XG30 and WT133 were transgressive segregants from XG and WT RIL populations with their transgressive segregation over their high parent at 1.19–4.80 % SPC; at the second cycle the transgressive lines derived from five superior families of the cross WT133 × XG30 performed 3.72–4.82 % SPC over their high parent WT133 and 8.22–9.32 % SPC over their first cycle high parent Wan 82-178 and TSBPHDJ. It demonstrated the potential of transgressive segregation through genetic recombination in breeding for SPC of soybean. The major genetic mechanism was explained by the three major QTL loci of *Prot-08-1*, *Prot-14-1* and *Prot-19-2*. It also demonstrated the potential of marker-assisted breeding in keeping the three major loci genetically controlled. However, there appeared some phenomena that could not be well explained with only the three major QTL, for example, the above maximums are even higher than the WX1-139-derived $A_2A_2B_2B_2L_2L_2$ lines, which suggests more QTL should be explored in the future studies.

Discussion

Potential and mechanism of recombination for transgressive breeding

In the present study, the SPC of the four original parents was 35.35–44.83 %, after two cycles of recombination the genetic progress of 9.32 % was obtained. It showed the great potential of recombination for transgressive breeding.

One of the reasons is the parental materials hold complementary genetic structure based on a high SPC level. In the present case, three of the four parents have their SPC about 45 % and the genetic structure on the three major loci is completely complementary to each other for XG30 ($A_2A_2B_2B_2L_1L_1$) and WT133 ($A_1A_1B_1B_1L_2L_2$). The other

Table 7 Comparisons of seed protein content among the three major QTL genotypes in WX population

Positive locus number	Genotype	WX1								WX2			
		$F_{2:3}$				$F_{2:4}$				$F_{2:3}$			
		N	Mean	Min	Max	*N*	Mean	Min	Max	*N*	Mean	Min	Max
Three	$A_2\_B_2\_L_2\_$	82	47.89 a	43.90	51.20	81	48.33 a	44.40	51.60	109	47.16 ab	42.40	52.20
Two	$A_2\_B_2\_L_1L_1$	37	47.75 ab	44.65	51.30	34	47.83 ab	45.85	51.30	25	47.61 a	42.90	51.20
	$A_2\_B_1B_1L_2\_$	34	47.90 a	45.25	50.30	31	47.65 ab	45.25	51.35	32	47.17 ab	44.10	50.10
	$A_1A_1B_2\_L_2\_$	28	46.64 bc	44.65	48.95	29	46.92 bcd	43.80	49.20	28	46.73 ab	42.80	50.50
One	$A_2\_B_1B_1L_1L_1$	13	46.08 c	43.85	48.15	11	46.36 cd	43.65	49.05	10	46.16 ab	42.50	50.30
	$A_1A_1B_2\_L_1L_1$	10	47.02 abc	45.40	49.05	8	47.19 abc	44.95	49.00	5	45.48 bc	44.10	47.10
	$A_1A_1B_1B_1L_2\_$	7	45.89c	44.60	47.70	5	45.76 d	43.95	47.75	14	45.48 bc	44.10	47.10
Zero	$A_1A_1B_1B_1L_1L_1$	5	46.19c	44.40	48.50	6	46.30 cd	44.80	48.85	4	44.03 c	40.50	46.40
	Total	216	47.45	43.85	51.30	205	47.67	43.65	51.60	227	46.85	40.50	52.20

In Genotype column A, B and L represent the QTL of *Prot-08-1, Prot-14-1* and *Prot-19-2,* respectively, and the subscribe 1 means negative allele and subscribe 2 means positive allele. Accordingly, the genotype of WT133 is $A_1A_1B_1B_1L_2L_2$, that of XG30 is $A_2A_2B_2B_2L_1L_1$ and that of Xianjin 2, Gantai-2-2, Wan 82-178 and TSBPHDJ is $A_1A_1B_2B_2L_1L_1$, $A_2A_2B_1B_1L_2L_2$, $A_1A_1B_1B_1L_2L_2$ and $A_1A_1B_1B_1L_2L_2$, respectively. In the WX population, the additive effect of *Prot-08-1*, *Prot-14-1* and *Prot-19-2* were 1.11, 0.45 and 0.31 %, while the partial dominance effect of *Prot-08-1*, *Prot-14-1* and *Prot-19-2* were −0.31, 0.24 and −0.10 %, indicating A_1, B_2 and L_1 were dominant over A_2, B_1 and L_2, respectively

The additive effect and dominant effect were estimated from $1/2(X_2X_2\text{-}X_1X_1)$ and $(X_1X_2\text{-}X_1X_1)\text{-}\ 1/2(X_2X_2\text{-}X_1X_1)$, respectively, where *X* represents an allele of a QTL

Note that in WX population, the additive effect of *Prot-14-1* is positive while that in NJRSXG is negative. There is no conflict between the two estimates because the P_1 genotype (Xianjin 2) contains a positive allele B_2 which causes the estimate of allele effect negative in linkage mapping since in mapping procedure the model is set as P_2 with larger effect allele. Anyway, the size of the estimates of allele effect in NJRSXG and WX is not the same since it is only relative to the respective populations

In the column of Mean, the letters indicate significance at $p < 0.05$ according to Duncan's new multiple range test

reason is the major QTL in the genetic system of the four parents were detected from which marker-assisted selection could be acted on.

Accordingly, the recombination breeding for transgressive segregation should be of more potential if the genetic constitution of the parental materials can be thoroughly explored. In fact, in the present case, only five additive QTL with 32.16 % contribution to PV and two epistatic QTL pairs with 2.96 % contribution to PV were detected and only three of the five major QTL were used in marker-assisted selection, the remnant 53.99 % contribution to PV was due to the undetected collective of minor QTL (or missing heritability) which was neglected in marker-assisted selection. This kind of situation is not rare according to Gai et al. (2012), who reported that there existed different constitutions of major, small and UCM QTL among the 110 data sets of 81 agronomically important traits in soybean, and sometimes UCM QTL accounted for more than 50 % of the genetic variation in a number of traits. Kim et al. (2014) indicated the improvement of phenotype precision, enlargement of population size and increase of marker density may change some UCM QTL into detected minor QTL. However, linkage mapping with bi-parent RIL population could detect only two alleles on a locus while association mapping which could take the germplasm resources as mapping population was considered as a powerful procedure to detect more QTL and alleles per locus (Yan et al. 2011; Hwang et al. 2014). It is sure that if the genetic constitution of the parental materials could be explored thoroughly, the breeding efficiency for transgressive segregation would be further improved.

Potential of marker-assisted selection for transgressive breeding

In the present study, three major additive QTL were used in marker-assisted breeding, with which the transgressive SPC were achieved successfully. The result evidenced the usefulness of the marker-assisted improvement of SPC based on major additive QTL. In the literature, Sebolt et al. (2000) also used a QTL allele from wild soybean to increase SPC in soybean.

However, the three major additive QTL explained only a small part (31.69 %) of the total genetic variation. What is the reason for that the present case has provided a successful transgressive progress in the improvement of SPC? In addition to marker-assisted selection on the three major QTL, one important reason might be the marker-assisted selection in fact integrated with the phenotypic selection since all the materials were evaluated for their SPC and the superior lines were confirmed accordingly; while another

Table 8 Transgressive segregation of seed protein content in families derived from WX1

Family	WT133	XG30	$F_{2:3}$	$F_{2:4}$	$F_{2:5:6}$										
					Class mid-point value								Mean	Min	Max	SD	CV
					<48	48.5	49.5	50.5	51.5	52.5	53.5	≥54	∑f				
WX1-139 ($A_1A_2B_2B_2L_2L_2$)	49.33	46.32	49.35	49.25	14	12	7	12	6	2	1		54	49.22	44.30	53.05	1.77	3.60
WX1-158 ($A_1A_2B_1B_2L_1L_2$)	49.33	46.32	49.85	51.20		5	3	9	24	16	3		60	51.40	48.05	53.35	1.20	2.33
WX1-180 ($A_1A_2B_1B_2L_1L_2$)	49.33	46.32	49.50	49.35	3	8	14	8	4	9	4	1	51	50.48	47.15	54.15	1.77	3.51
WX1-195 ($A_1A_2B_2B_2L_1L_1$)	49.33	46.32	49.65	49.10		6	19	11	8	6	1	1	52	50.46	48.55	54.05	1.33	2.64
WX1-205 ($A_1A_2B_1B_2L_1L_1$)	49.33	46.32	51.30	49.80	2	3	4	7	21	10	4	1	52	51.20	47.65	54.10	1.48	2.89
Overall	49.33	46.32	49.93	49.74	19	34	47	47	63	43	13	3	269	50.55	44.30	54.15	1.71	3.38

In parentheses is the genotype of the family's ancestor F_2 plant; the genotypes in the listed derived generations ($F_{2:3}$, $F_{2:4}$ and $F_{2:5:6}$) are mixtures resulted from further segregation

The SPC of WT133 and XG30 in the second cycle experiment is 0.94 and 0.79 % more than that in the first cycle experiment, respectively, in addition to the correction for the instruments

SD standard deviation, *CV* coefficient of variation

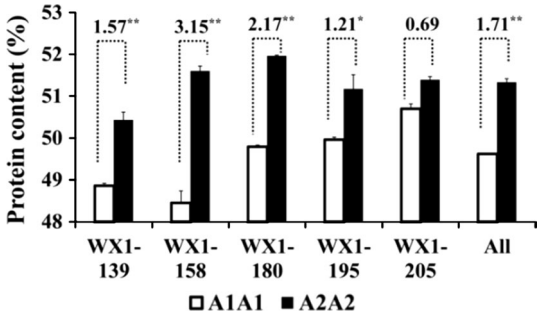

Fig. 4 The average difference of protein content between the two homozygotes on *Prot-08-1* in families derived from WX1. The number above the histogram is the average difference between the two homozygotes on *Prot-08-1*. The phenotypic variation among the two homozygotes of *Prot-08-1* were 15.31, 26.78, 22.23, 14.64, 4.35 and 24.11 % for WX1-139, WX1-158, WX1-180, WX1-195, WX1-205 and all, respectively. *Asterisk* and *double asterisk* represent significant at $p < 0.05$ and $p < 0.01$, respectively

reason might be the complementary recombination on other QTL undetected in the mapping process. For example, at the first cycle, WT133 with SPC 4.80 % over its high parent was selected from Wan 82-178×TSBPHDJ, but the genotypes on the three major QTL loci, even WT133 itself were the same as $A_1A_1B_1B_1L_2L_2$. Thus the 4.80 % SPC transgressive amount at the first cycle and 9.32 % after the second cycle might be due to recombination on other loci, such as possible *X* and *Y* loci in Table 5 and even many undetected loci. The phenotypic selection helped in the improvement of genotypes of small or minor QTL in addition to the three major QTL. Therefore, marker-assisted selection integrated with phenotypic selection should be a useful strategy in the improvement of traits easy to be evaluated but with unenough mapping information.

Anyway, if a fine genetic dissection of the materials can be achieved or the genome-wide QTL can be detected, marker-assisted breeding, including marker-assisted cross design and marker-assisted progeny selection will be highly efficient. Gai et al. (2012) proposed to establish genome-wide QTL-allele matrix for marker-assisted breeding. It considered that the QTL system for a trait could be dissected using GWAS procedure, then the detected QTL along with their multiple alleles could be used to establish QTL-allele matrix for all the breeding materials which can be used for marker-assisted cross design. It should have potential for transgressive breeding through choosing parents genetically complementary to each other.

Author contribution statement YHZ, TJZ and JYG designed the study and drafted the report; YHZ, MFL and GNX conducted the field experiment; YHZ, JBH, YFW, SPY and YL conducted the lab work and molecular data analysis.

Acknowledgments This research was supported by the National Key Basic Research Program of China (2011CB1093), the National Hightech R&D Program of China (2011AA10A105, 2012AA101106), the MOE 111 Project (B08025), the MOE Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT13073), the MOA Public Profit Program (201203026-4), the Jiangsu Higher Education PAPD Program and the Jiangsu JCIC-MCP program.

Conflict of interest The authors have declared that no competing or conflicts of interest exist.

References

Chen QS, Zhang ZC, Liu CY, Xin DW, Qiu HM, Shan DP, Shan CY, Hu GH (2007) QTL Analysis of Major Agronomic Traits in Soybean. Agr Sci China 6:399–405

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos T Roy Soc B 363:557–572

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

Gai JY, Chen L, Zhang YH, Zhao TJ, Xing GN, Xing H (2012) Genome-wide genetic dissection of germplasm resources and implications for breeding by design in soybean. Breed Sci 61:495–510

Hanson C, Robinson H, Comstock R (1956) Biometrical studies of yield in segregating populations of Korean Lespedeza. Agronomy J 48:268–272

Hwang EY, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genom 15:1

Kim H, Xing G, Wang Y, Zhao T, Yu D, Yang S, Li Y, Chen S, Palmer R, Gai J (2014) Constitution of resistance to common cutworm in terms of antibiosis and antixenosis in soybean RIL populations. Euphytica 196:137–154

Korir PC, Qi B, Wang Y, Zhao T, Yu D, Chen S, Gai J (2011) A study on relative importance of additive, epistasis and unmapped QTL for Aluminium tolerance at seedling stage in soybean. Plant Breed 130:551–562

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.]. Agr Sci China 9:1108–1116

Liu SH, Zhou RB, Gai JY (2009) A comparative analysis of protein and fat content between wild and cultivated soybean in China. Soybean Sci 28:566–573

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321–4325

Orf JH, Chase K, Jarvik T, Mansur LM, Cregan PB, Adler FR, Lark KG (1999) Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Sci 39:1642–1651

Peleman JD, van der Voort JR (2003) Breeding by design. Trend Plant Sci 8:330–334

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Sebastian SA, Streit LG, Stephens PA, Thompson JA, Hedges BR, Fabrizius MA, Soper JF, Schmidt DH, Kallem RL, Hinds MA, Feng L, Hoeck JA (2010) Context-specific marker-assisted selection for improved grain yield in elite soybean populations. Crop Sci 50:1196–1206

Sebolt AM, Shoemaker RC, Diers BW (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40:1438–1444

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE, Cregan PB (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet 109:122–128

Song Q, Jia G, Zhu Y, Grant D, Nelson RT, Hwang EY, Hyten DL, Cregan PB (2010) Abundance of SSR Motifs and Development of Candidate Polymorphic SSR Markers (BARCSOYSSR_1.0) in Soybean. Crop Sci 50:1950–1960

Su CF, Zhao TJ, Gai JY (2010) Simulation comparisons of effectiveness among QTL mapping procedures of different statistical genetic models. Acta Agron Sinica 36:1100–1107

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

Wang YF (2009) Genomic characterization of simple sequence repeats and establishment, integration and application of high density genetic linkage map in soybean. Ph.D. dissertation, Nanjing Agricultural University, Nanjing, People's Republic of China

Wang HW (2011) Genetic dissection and elite allele identification of seed traits in soybean culivars released from Huang-huai valleys and southern China. Master dissertation, Nanjing Agricultural University, Nanjing, People's Republic of China

Wang YF, Lu JJ, Chen SY, Shu LP, Palmer RG, Xing GN, Li Y, Yang SP, Yu DY, Zhao TJ, Gai JY (2014) Exploration of presence/absence variation and corresponding polymorphic markers in soybean genome. J Integrat Plant Biol 56:1009–1019

Xing GN, Zhou B, Wang YF, Zhao TJ, Yu DY, Chen SY, Gai JY (2012) Genetic components and major QTL confer resistance to bean pyralid (*Lamprosema indicata* Fabricius) under multiple environments in four RIL populations of soybean. Theor Appl Genet 125:859–875

Yan JB, Warburton M, Crouch J (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. Crop Sci 51:433–449

Yang J, Hu C, Hu H, Yu R, Xia Z, Ye X, Zhu J (2008) QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. Bioinformatics 24:721–723