8-9-2017

Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits

Lillian F. Brzostowski
University of Illinois, brzosto2@illinois.edu

Timothy I. Pruski
Bayer CropScience, White Heath, IL

James Specht
University of Nebraska - Lincoln, jspecht1@unl.edu

Brian W. Diers
University of Illinois, bdiers@illinois.edu

Published in *Theoretical* and *Applied Genetics* 130 (2017), pp. 2315–2326.
DOI 10.1007/s00122-017-2961-x

Submitted 31 March 2017; accepted 31 July 2017; published 9 August 2017

Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits

Lillian F. Brzostowski,[1] Timothy I. Pruski,[2] James E. Specht,[3] and Brian W. Diers[1]

1 Department of Crop Sciences, University of Illinois, 1101 W. Peabody Drive, Urbana, IL 61801, USA
2 Bayer CropScience, 21 County Road 1200 North, White Heath, IL 61884, USA
3 Department of Agronomy and Horticulture, University of Nebraska, 363 Keim Hall, Lincoln, NE 68583, USA

Corresponding author — Brian W. Diers bdiers@illinois.edu

Abstract

***Key message* — Evaluation of seed protein alleles in soybean populations showed that an increase in protein concentration is generally associated with a decrease in oil concentration and yield.**

Abstract — Soybean [*Glycine max* (L.) Merrill] meal is one of the most important plant-based protein sources in the world. Developing cultivars high in seed protein concentration and seed yield is a difficult task because the traits have an inverse relationship. Over two decades ago, a protein quantitative trait loci (QTL) was mapped on chromosome (chr) 20, and this QTL has been mapped to the same position in several studies and given the confirmed QTL designation cqSeed protein-003. In addition, the *wp* allele on chr 2, which confers pink flower color, has also been associated with increased protein concentration. The objective of our study was to evaluate the effect of cqSeed protein-003 and the *wp* locus on seed composition and agronomic traits in elite soybean backgrounds adapted to the Midwestern USA. Segregating populations of isogenic lines were developed to test the *wp* allele and the chr 20 high protein QTL alleles from Danbaekkong (PI619083) and *Glycine soja* PI468916 at cqSeed protein-003. An increase in protein concentration and decrease in yield were generally coupled with the high protein alleles at cqSeed protein-003 across populations, whereas the effects of *wp* on protein concentration and yield were variable. These results not only demonstrate the difficulty in developing cultivars with increased protein and yield but also provide information for breeding programs seeking to improve seed composition and agronomic traits simultaneously.

Introduction

Soybean is grown as a source of protein and oil, and soybean seed averages approximately 350 g kg^{-1} protein (130 g kg^{-1} moisture basis). The seed contains a good balance of the amino acids necessary to meet the dietary requirements of swine and poultry (Liu 1997; Kerley and Allee 2003; Cromwell 2012), which makes it an exceptional source of protein meal for livestock and a leading source of plant-based protein in the world (Wilson 2008; Cromwell 2012). It is important for breeders to continue to develop soybean cultivars that maintain and improve current protein levels, so soybean will continue its prominence as a livestock feed.

There is considerable range in seed-protein concentration in soybean germplasm accessions. In the USDA Soybean Germplasm Collection, there are accessions with protein concentrations, on a 130 g kg^{-1} moisture basis, as low as 276 g kg^{-1} and as high as 504 g kg^{-1} (USDA 2017). Additionally, protein concentration is a highly heritable trait with reported heritabilities of up to 0.99 (Brummer et al. 1997; Chung et al. 2003; Eskandari et al. 2013).

Although a high heritability and a substantive range in genotypic values should make increasing seed-protein concentration an obtainable objective for breeding programs, complex relationships between protein, oil, and yield have made it difficult to efficiently combine high values for each

of these three desirable traits into a single cultivar. The negative correlations between protein and oil concentration and protein and yield have been well established (Hartwig and Kilen 1991; Wilcox and Cavins 1995; Sebolt et al. 2000; Wilcox and Shibles 2001; Chung et al. 2003; Wilson 2004; Eskandari et al. 2013; Bandillo et al. 2015). A leading hypothesis for the negative correlations is the physiological relationship between nitrogen and carbon supply. Both nitrogen and carbon supply play a role in yield and seed composition and are affected by nitrogen accumulation, partitioning, and whole plant remobilization (Sinclair and de Wit 1975). Protein and oil rely on the same limited carbon energy supply, and each component has a different energy requirement (Hanson et al. 1961; Shimura and Hanson 1970; Chung et al. 2003). On a genetic basis, hypotheses for the negative genetic correlations between protein and yield and protein and oil include but are not limited to pleiotropic effects or linkage drag (Chung et al. 2003; Nichols et al. 2006; Bandillo et al. 2015).

While the negative correlation between yield and protein concentration is strong, it is weaker than that between protein and oil (Chung et al. 2003; Cober and Voldeng 2000). There is evidence that the relationship between yield and protein can be modulated. Individual lines and groups of lines with elevated protein and yield have been identified in studies in which a high protein phenotype present in a low yielding parent has been backcrossed into a low protein, high yielding parent (Wehrmann et al. 1987; Wilcox and Cavins 1995). In addition, recurrent selection and restricted index selection have been successfully used in developing high protein and high yielding lines (Brim and Burton 1979; Holbrook et al. 1989). Following 26 generations of random mating to reduce linkage disequilibrium, Recker et al. (2014) observed no significant genetic correlations between yield and protein. However, a significant negative correlation between oil and protein was still observed, which provides further evidence of a pleiotropic relationship between the two traits. The aforementioned studies suggest there can be success in increasing yield and protein simultaneously when the appropriate breeding strategy is implemented. Evaluation and characterization of QTL associated with protein concentration can provide valuable information to help determine the best breeding scheme to meet seed composition and yield objectives.

Quantitative trait loci (QTL) for protein concentration have been mapped to all soybean chromosomes (Soybase 2017). In one of the first QTL mapping studies in soybean, Diers et al. (1992) mapped two seed protein QTL in a population derived from a cross between the *G. max* experimental line, A81-356022, and the *Glycine soja* Siebold and Zucc. plant introduction, PI468916. One protein QTL was mapped to chromosome (chr) 15 [formerly linkage group (LG) E], whereas the other mapped to chr 20 (formerly LG I). These QTL were confirmed based on guidelines set forth by the Soybean Genetics Committee (Soybase 2017). The QTL on chr 15 was given the designation cqProt-001 (Fasoula et al. 2004), and the QTL on chr 20 was designated cqProt-003 (Nichols et al. 2006). These designations have since been updated on the Soybase website (Soybase 2017) and are now listed as cqSeed protein-001 and cqSeed protein-003, respectively. Protein QTL have been mapped from several sources to the same genomic regions on chr 15 (Lee et al. 1996; Brummer et al. 1997; Fasoula et al. 2004; Kim et al. 2015; Phansak et al. 2016) and chr 20 (Brummer et al. 1997; Sebolt et al. 2000; Chung et al. 2003; Wang et al. 2014; Warrington et al. 2015; Phansak et al. 2016) suggesting these loci may have several alleles or the same alleles may be in several accessions or alternatively, there could be multiple closely linked QTL in these intervals. Follow-up studies have sought to refine the locations of the chr 15 and chr 20 QTL using advanced genetics techniques (Bolon et al. 2010; Hwang et al. 2014; Vaughn et al. 2014; Kim et al. 2015; Bandillo et al. 2015).

The Korean cultivar Danbaekkong (PI619083) contains a high protein allele at the chr 20 QTL (Harris 2001; Yates 2006; Warrington et al. 2015). Danbaekkong is a late maturity group (MG) IV soyfood cultivar (Kim et al. 1996). Although PI468916 and Danbaekkong have protein QTL that map to the same region on chr 20, it is unknown whether their alleles are the same or different. We will herein refer to the high protein QTL allele from Danbaekkong as CHR20-D and the high protein QTL allele from PI468916 as CHR20-PI.

CHR20-PI has been evaluated across northern US soybean backgrounds and was found to be associated with increased protein, reduced yield, reduced oil, smaller seeds, taller plants, and/or earlier maturity (Sebolt et al. 2000; Nichols et al. 2006). Evaluation of CHR20-D in southern US backgrounds and locations has shown an inconsistent association with yield, and it has been suggested that the Danbaekkong high protein allele could be successfully used to develop lines with high protein and yield (Harris 2001; Yates 2006). A recently released MGIII cultivar containing CHR20-D was demonstrated to have increased protein concentration and no yield loss compared to the checks (Mian et al. 2017). However, CHR20-D has not been directly evaluated in northern US soybean germplasm, and thus, there is a need to determine whether its effect on protein and other traits is similar to what was observed by the high protein allele for CHR20-PI.

The recessive *wp* allele, which confers pink flower color, was mapped to chr 2 (formerly LG D1b), and this allele was found to be associated with increased seed protein concentration (Stephens and Nickell 1992; Stephens et al. 1993). Stephens et al. (1993) also showed that the *wp* allele was

associated with larger seeds and decreased seed oil concentration (Stephens et al. 1993). Hegstad et al. (2000) observed lines containing the *wp* allele in two populations had increased protein concentration, decreased oil concentration, later maturity, and increased plant height. Additionally, significant yield reductions associated with *wp* were observed in one population. Zabala and Vodkin (2005) determined that the pink flower color caused by the *wp* allele was the result of the insertion of a transposable element in the flavanone 3-hydroxylase gene 1. To date, the *wp* allele has not been tested in a background other than the one in which it was first discovered. Before a protein-increasing QTL or gene can be widely used in breeding programs, it is important to analyze its effect, not only on protein concentration, but also agronomic traits, especially yield, in various high-yield genetic backgrounds. The objective of this study is to test the effect of CHR20-D, CHR20-PI, and *wp* on protein concentration and other agronomic traits in multiple genetic backgrounds.

Materials and methods

Plant material

Population development CHR20–D

Two populations of isogenic lines segregating for CHR20-D were developed. The donor parent Danbaekkong was mated to the recurrent parents 'Dwight', a late MG II cultivar (Nickell et al. 1998), and LD02-5025, a late MG II elite breeding line (Cary and Diers 2007). An F_2 plant that was homozygous for CHR20-D was selected from each mating and backcrossed to the respective recurrent parent using simple sequence repeat markers (SSR) linked to the chr 20 QTL to facilitate the introgression without the need to analyze the seed protein contents of backcross progeny. An additional generation of backcrossing was conducted to reach the backcross-two F_1 (BC_2F_1) generation. After each generation of backcrossing, the presence of the CHR20-D allele was verified using several SSR markers linked to the QTL including Satt614, Satt239, and Satt354 (Nichols et al. 2006). During backcrossing, no background selection was done to increase the recovery of the recurrent parent. Heterozygous BC_2F_1 plants were selfed to produce BC_2F_2 seed. Plants in this selfed generation were genotyped with markers linked to the QTL to identify plants homozygous for (1) the high protein allele (i.e., CHR20-D) from the donor and (2) the corresponding low-protein allele from the recurrent parent. Any lines exhibiting a recombination between the SSR donor and recurrent parent markers were discarded. Two populations of BC_2F_2-derived lines, one for each recurrent parent, plus their respective recurrent parents and check cultivars were grown in the field in 2013 and 2014. There were 39 lines in the LD00-5025 population (17 homozygous for CHR20-D at markers linked to the chr 20 QTL and 22 homozygous for the low protein allele at markers linked to the chr 20 QTL) and 47 lines in the Dwight population (24 homozygous for CHR20-D at markers linked to the QTL, 23 homozygous for the low protein allele at markers linked to the chr 20 QTL).

Population development CHR20-PI and wp

Four populations were developed from four separate backcrosses (BC_4) in which one of four different Illinois-adapted genotypes were used as a recurrent parent. These parents included the two maturity group II cultivars Dwight (Nickell et al. 1998) and Loda (Nickell et al. 2001), and the two maturity group IV experimental lines LS93-0375 (Schmidt and Klein 2002) and C1981 (Nowling 2001). The donor parent possessing the high protein CHR20-PI allele originated from a BC_3F_4 population (A81-356022 (4) × PI 468916) described by Sebolt et al. (2000). The donor parent for the *wp* allele was a F_4-derived line from the cross of two parents with pink flowers, LN89-5320 × LN89-5322 (Stephens and Nickell 1992; Stephens et al. 1993). The high protein QTL alleles in the two donor parents were introgressed into each of the four genetic backgrounds in the following manner. Presence of CHR20-PI was verified in BC_nF_1 plants with the SSR markers Satt239 and Satt496 (Nichols et al. 2006). Lines with a recombination between the two markers were discarded, and selected BC_nF_1 plants containing CHR20-PI were then mated to the recurrent parents. The presence of the *wp* allele was verified by performing progeny tests (i.e., progenies fixed for green hypocotyl color-inferred parent was homozygous for *wp*, progenies fixed for purple hypocotyl color-inferred parent was homozygous for *Wp*, etc.) with the BC_nF_2 seed and occurred after the next backcross (BC_{n+1}) had taken place. Progeny tests from the previous generation were used to identify the $BC_{n+1}F_1$ seed to be genotyped with molecular markers to verify the presence of the CHR20-PI allele from PI468916. BC_4F_1 plants predicted to be heterozygous for alleles at both QTLs within each background were selfed, and single-seed descent was performed to produce BC_4F_4 seed. BC_4F_4 plants homozygous in both QTLs were selected and selfed to form populations of BC_4F_4-derived lines. Molecular markers linked to CHR20-PI and progeny tests for the *wp* locus were used to assess the genotype of the lines, and lines with recombination between markers in the region were discarded. No background selection was done based on phenotypes or with markers during backcrossing.

Field tests

Environments and check cultivars CHR20-D populations

In 2013 and 2014, populations of BC_2F_2–derived lines were evaluated at the Crop Sciences Research and Education Center in Urbana, IL and in a grower's field near Pontiac, IL. Planting dates were as follows: Pontiac, IL 2013, May 14; Urbana, IL 2013, May 15; Pontiac, IL 2014, May 7; Urbana, IL 2014, May 21. The check cultivar was IA2102 (Crochet and Hughes 2011) for both populations.

Environments and check cultivars CHR20-PI and wp populations

MG II BC_4F_4 populations Maturity group II BC_4F_4 populations were grown at the Northern Illinois Agronomy Research Center in DeKalb, IL in 2008, the Crop Sciences Research and Education Center in Urbana, IL in 2007 and 2008, a Mead, NE rain-fed (Rf) location in 2007, and a Mead, NE irrigated (Ir) location in 2007 for a total of five environments. Planting dates were as follows: Urbana, IL 2007, May 16; Mead Rf, NE 2007 and Mead Ir, NE 2007, May 17; DeKalb, IL 2008, May 20; Urbana, IL 2008, May 28. Check cultivars were LD02-4485 (Abney and Crochet 2006) and IA2068 (Abney and Crochet 2003) in the Loda backcross population, whereas the Dwight backcross population included only LD02-4485. The respective recurrent parent for each population was also included in the trials. There were 65 lines in the Loda population and 71 lines in the Dwight population.

MGIV BC_4F_4 populations Maturity group IV BC_4F_4 populations were planted at the Crop Sciences Research and Education Center in Urbana, IL during 2007 and 2008, a Mead, NE Rf location in 2007, and a Mead, NE Ir location in 2007 for a total of four environments. Planting dates were the same as those previously mentioned for the MG II populations. LD00-3309 (Diers et al. 2006) was a check cultivar in both MG IV populations while the LS93-0375 population included the cultivar Macon (Nickell et al. 1996) as an additional check. The recurrent parent for each population was also included in the field evaluations. There were 75 lines in the LS93-0375 population and 49 lines in the C1981 population.

Field evaluation and phenotypic measurements for all populations

Populations were blocked separately, and the lines plus the recurrent parents and check cultivars were arranged in a randomized complete block design. The CHR20-D populations were grown in non-replicated tests, and the CHR20-PI populations were replicated twice. All populations were planted in two-row plots, 3.6 m long using a four-row ALMACO plot planter (ALMACO Iowa). Row spacing was 0.76 m, and seeding rate was ~27 seeds per meter. All environments were rain-fed with the exception of Mead, NE (Ir). Plots were rated for maturity date, plant height, and lodging. Plant height was measured in cm as the distance between the soil surface and the top node on the main stem. Maturity was the date when 95% of the pods reached mature color (R8 described by Fehr et al. 1971) with September 1 recorded as 901. Lodging was rated on a scale of 1 and 5, with 1 equaling all plants erect and 5 equaling all plants prostrate. Seed yield was measured at maturity using an ALMACO plot combine, adjusted to 130 g kg^{-1} moisture, and reported as kg ha^{-1}. Additionally, a Perten DA 7250 NIR analyzer was used to determine protein and oil concentration on a 130 g kg^{-1} moisture basis for the CHR20-D populations (Perten Hagersten Sweden). This analysis was conducted with whole seed samples from each plot using the factory calibration and each sample was analyzed two times and the average recorded. Seed protein and oil concentration analysis for the CHR20-PI and *wp* locus populations was performed at the USDA Northern Regional Research Center in Peoria, IL on whole seed with near infrared transmittance and also reported on a 130 g kg^{-1} moisture basis.

DNA extraction and genetic marker analysis for all populations

Genomic DNA was isolated from young trifoliolate leaves by a modified CTAB method described by Keim et al. (1988) or a quick DNA extraction method described by Bell-Johnson et al. (1998). Polymorphic SSR markers were used to perform polymerase chain reactions according to Cregan and Quigley (1997). Amplification products were separated in 6% (w/v) non-denaturing polyacrylamide gels by electrophoresis (Wang et al. 2003).

Statistical analysis for all populations

All data were subjected to analysis of variance using SAS v9.4 (SAS Institute 2016) PROC MIXED. Data were analyzed across and within locations and an environment was a year by location combination. Marker genotype and lines nested within marker genotype were considered to be fixed effects, whereas replicate and environment were treated as random effects. Degrees of freedom were calculated according to the Kenward–Roger method (Littell et al. 2006).

Results

Chr20-D

CHR20-D was evaluated in the LD02-5025 and Dwight backgrounds, and each population was evaluated for seed composition and agronomic traits at four environments in Illinois. Although these populations were developed through only two backcrosses, the lines in each population were phenotypically similar to the recurrent parents. The population means for maturity were within 2 days of the recurrent parents and the mean plant height was within 2 cm of the recurrent parents. Because the effects of the CHR20-D were estimated using markers linked to this QTL, residual alleles from Danbaekkong segregating in the populations would have likely had a minimal impact on the estimated effects of CHR20-D.

For both backgrounds, marker alleles from Danbaekkong linked to CHR20-D were associated with a significant ($P < 0.05$) increase in protein concentration, decreased oil concentration, and increased lodging score compared to the recurrent parent allele across environments (Tables 1, 2). In addition, lines containing CHR20-D had a significant ($P < 0.0001$) yield reduction across environments compared to lines containing the recurrent parent allele for both backgrounds (Table 1). This difference was −455 kg ha^{-1} in the LD02-5025 background and −363 kg ha^{-1} in the Dwight background, which represent a seed yield decrease associated with the introgression of the donor parent high protein allele. Maturity date was not significant ($P < 0.05$) over environments in the LD02-5025 population, but was significant for the Dwight population with lines containing the Dwight allele maturing 2 days earlier than lines with the Danbaekkong allele (Table 2). Additionally, a significant marker genotype × environment interaction was observed for protein and oil concentration in both populations. The marker genotype × environment interactions for yield were non-significant.

For the LD02-5025 population, the lines containing the high protein QTL allele had increased average protein concentration and decreased oil concentration for each environment with the exception of Pontiac in 2013 (Table 1). These significant differences ranged from 25 to 31 g kg^{-1} for protein concentration and −10 to −14 g kg^{-1} for oil concentration. Within all four environments, lines with the high protein QTL allele on average yielded significantly ($P < 0.05$) less than lines with the LD02-5025 allele, and this difference ranged from −273 to −558 kg ha^{-1}.

Similar trends were observed in the Dwight population within environments for protein concentration, oil concentration, and yield (Table 1). Lines with the high protein QTL allele had significantly increased average protein concentration and decreased oil concentration compared to lines with the Dwight allele in the Urbana 2013, Urbana 2014, and Pontiac 2014 environments. These significant differences ranged from a 19 to 28 g kg^{-1} increase in protein concentration and a coupled −7 to −14 g kg^{-1} decrease in oil concentration. In addition, lines with CHR20-D yielded significantly less than those with the Dwight allele in all four environments. The observed difference ranged from −239 at Pontiac 2014 to −496 kg ha^{-1} at Urbana 2014.

CHR20-PI and *wp*

Four populations were developed via backcrossing to test the effect of CHR20-PI and *wp* on seed composition and agronomic traits. The Loda and Dwight populations were evaluated in five environments while the LS93-0375 and C1981 populations were evaluated in four environments. The population mean across environments was within 1 day of the recurrent parent for the Loda population, 3 days for the Dwight population, 5 days for the LS93-0375 population, and 6 days for the C1981 population. Across environments, a significant marker genotype × environment interaction was detected for protein within all backgrounds. Additional significant marker genotype × environment interactions were population-specific. Within and across environments, CHR20-PI was associated with significantly increased protein concentration and decreased oil concentration compared to the recurrent parent allele for all four populations (Table 3). The magnitude of the effect was dependent upon genetic background and environment. Within and across environments, the associated effect of CHR20-PI on yield was variable, although that variability did not include an example of a significant yield increase. Across environments, lines containing CHR20-PI had significantly reduced yields in the Dwight and C1981 populations, but such lines in the Loda and LS93-0375 populations did not exhibit significant yield depression. Within each population, CHR20- PI was significantly associated with a decreased maturity date of 1–3 days across environments (Table 2). A significant increase in plant height was also observed across environments in the Loda, Dwight, and LS93-0375 populations with plants containing the donor allele averaging 2.1–3.3 cm taller than those containing the recurrent allele. Significant associations were not observed for lodging in any of the four populations. While CHR20-PI was consistently associated with an increase in protein concentration and a decrease in oil concentration, the *wp* allele had a non-significant effect on oil concentration and a variable effect on protein concentration across environments when lines homozygous for *wp* were compared to lines containing no high protein alleles (Table 4). Across environments, the

Table 1. The impact on seed yield, protein, and oil (130 g kg^{-1} moisture basis) when the protein-increasing allele of the chr 20 QTL from Danbaekkong (CHR20-D) was introgressed into the LD02-5025 and Dwight backgrounds

Genetic background[a]	Environment[b]	Seed yield (kg ha^{-1})				Protein (g kg^{-1})				Oil (g kg^{-1})			
		Donor allele[c]	Recurrent allele[d]	Diff[e]	*P* value	Donor allele	Recurrent allele	Diff	*P* value	Donor allele	Recurrent allele	Diff	*P* value
LD02-5025	Pontiac 2013	4188	4650	−462	<0.0001	338	342	−4	ns	182	181	1	ns
	Urbana 2013	3741	4299	−558	<0.0001	374	343	31	<0.0001	170	184	−14	<0.0001
	Pontiac 2014	4591	5116	−525	<0.0001	373	348	25	<0.0001	166	176	−10	<0.0001
	Urbana 2014	3889	4162	−273	0.0047	367	337	30	<0.0001	169	181	−12	<0.0001
	Across	4102	4557	−455	<0.0001	363	343	20	<0.0001	172	181	−9	<0.0001
Dwight	Pontiac 2013	3598	3875	−277	0.001	350	350	0	ns	177	177	0	ns
	Urbana 2013	3339	3778	−439	<0.0001	378	350	28	<0.0001	171	185	−14	<0.0001
	Pontiac 2014	4149	4645	−496	<0.0001	369	346	23	<0.0001	171	179	−8	<0.0001
	Urbana 2014	3439	3678	−239	0.003	369	343	26	<0.0001	172	181	−9	<0.0001
	Across	3631	3994	−363	<0.0001	366	347	19	<0.0001	173	180	−7	<0.0001

ns non-significant
a. Recurrent parent of population
b. Location and year
c. Mean of lines predicted to be homozygous for the high-protein Danbaekkong allele at the chr 20 QTL (CHR20-D) based on the genetic markers Satt614, Satt239, and Satt354
d. Mean of lines predicted to be homozygous for the recurrent parent low-protein allele at the chr 20 QTL based on the genetic markers Satt614, Satt239, and Satt354
e. Difference between the means of lines that were homozygous for the donor and recurrent parent allele classes

Table 2. The impact on maturity, plant height, and plant lodging when the protein-increasing allele on chr 20 derived from Danbaekkong (CHR20-D) or derived from PI468916 (CHR20-PI) was introgressed into the listed genetic backgrounds

QTL allele[d]	Genetic background[e]	Maturity date[a]				Plant height (cm)[b]				Lodging (1–5)[c]			
		Donor allele[f]	Recurrent allele[g]	Diff[h]	*P* value	Donor allele	Recurrent allele	Diff	*P* value	Donor allele	Recurrent allele	Diff	*P* value
CHR20-D	LD02-5025	919	920	−1	ns	94.0	91.0	3.0	ns	2.7	2.2	0.5	<0.0001
	Dwight	918	920	−2	0.0016	94.3	93.2	1.1	ns	2.2	2.0	0.2	0.018
CHR20-PI	Loda	916	917	−1	0.0262	72.2	70.1	2.1	0.0174	2.1	2.0	0.1	ns
	Dwight	917	918	−1	0.0005	77.8	75.5	2.3	0.0004	1.7	1.5	0.2	ns
	LS93-0375	925	926	−1	0.0044	92.4	89.1	3.3	<0.0001	1.5	1.6	−0.1	ns
	C1981	928	931	−3	0.0058	106.4	104.4	2.0	ns	2.1	2.0	0.1	ns

ns non-significant

a. Characterized as the calendar date when 95% of pods have reached mature color (R8; Fehr et al. 1971) with September 1 equivalent to 901

b. Measured as the distance from the soil surface to the topmost node on the main stem

c. Lodging is visually rated on a 1 to 5 scale (i.e., 1 = all plants erect and 5 = all plants prostrate)

d. High protein allele name (that originated from one or the other donor parent—see caption)

e. Recurrent parent of population

f. Mean of lines predicted to be homozygous for the high protein CHR20-D allele based on the genetic markers Satt614, Satt239, and Satt354 or the high protein CHR20-PI allele based on the genetic markers Satt239 and Satt496

g. Mean of lines predicted to be homozygous for the chr 20 low protein allele of the recurrent parent based on the genetic markers Satt614, Satt239, and Satt354 or Satt239 and att496

h. Difference between the means of homozygous classes

Table 3. The impact on seed yield, protein, and oil (130 g kg−1 moisture basis) when the high protein QTL allele on chr 20 from PI468916 (CHR20-PI) was introgressed into the Loda, Dwight, LS93-0375, and C1981 backgrounds

		Seed yield (kg ha^{-1})				Protein (g kg^{-1})				Oil (g kg^{-1})			
Genetic background[a]	Environment[b]	Donor allele[c]	Recurrent allele[d]	Diff[e]	*P* value	Donor allele	Recurrent allele	Diff	*P* value	Donor allele	Recurrent allele	Diff	*P* value
Loda	Dekalb, IL 2008	3138	3188	−50	ns	356	344	12	<0.0001	174	181	−7	<0.0001
	Mead, NE (Ir) 2007	2819	3092	−273	0.0031	392	370	22	<0.0001	165	176	−11	<0.0001
	Mead, NE (Rf) 2007	2859	3115	−256	0.0036	383	361	22	<0.0001	171	181	−10	<0.0001
	Urbana, IL 2007	2510	2672	−162	0.0241	381	363	18	<0.0001	170	180	−10	<0.0001
	Urbana, IL 2008	2024	1953	71	ns	362	349	13	<0.0001	176	182	−6	<0.0001
	Across	2670	2804	−134	ns	375	357	18	<0.0001	171	180	−9	0.0004
Dwight	Dekalb, IL 2008	3659	3941	−282	<0.0001	354	342	12	<0.0001	165	174	−9	<0.0001
	Mead, NE (Ir) 2007	3542	4020	−478	<0.0001	386	360	26	<0.0001	153	166	−13	<0.0001
	Mead, NE (Rf) 2007	3486	3895	−409	<0.0001	384	358	26	<0.0001	159	171	−12	<0.0001
	Urbana, IL 2007	3233	3510	−277	<0.0001	364	347	17	<0.0001	162	172	−10	<0.0001
	Urbana, IL 2008	2853	2999	−146	ns	347	333	14	<0.0001	175	184	−9	<0.0001
	Across	3354	3673	−319	0.0053	367	348	19	0.0019	163	173	−10	<0.0001
LS93-0375	Mead, NE (Ir) 2007	4256	4513	−257	0.0003	398	373	24	<0.0001	156	169	−13	<0.0001
	Mead, NE (Rf) 2007	4133	4339	−206	0.0058	394	369	25	<0.0001	162	175	−13	<0.0001
	Urbana,IL 2007	3712	3861	−149	0.0483	386	365	21	<0.0001	164	177	−13	<0.0001
	Urbana, IL 2008	3240	3213	27	ns	374	358	16	<0.0001	174	183	−9	<0.0001
	Across	3835	3982	−147	ns	388	366	22	<0.0001	164	176	−12	<0.0001
C1981	Mead, NE (Ir) 2007	4003	4239	−236	ns	403	378	25	<0.0001	155	167	−12	<0.0001
	Mead, NE (Rf) 2007	3989	4370	−381	0.0010	404	374	30	<0.0001	157	171	−14	<0.0001
	Urbana, IL 2007	3424	3655	−231	0.0042	386	364	22	<0.0001	163	177	−14	<0.0001
	Urbana, IL 2008	3098	3330	−232	0.0008	375	360	15	<0.0001	168	179	−11	<0.0001
	Across	3629	3899	−270	0.0007	392	369	23	0.0005	161	174	−13	<0.0001

ns non-significant
a. Recurrent parent of population
b. Location and year
c. Mean of lines predicted to be homozygous for CHR20-PI based on the genetic markers Satt239 and Satt496
d. Mean of lines predicted to be homozygous for the recurrent parent allele at chr 20 based on the genetic markers Satt239 and Satt496
e. Difference between the means of homozygous classes

Table 4. Across environment means for seed yield, protein, oil, maturity, lodging and height of the lines homozygous for the lower protein recurrent parent allele on chr 20 and *Wp* on chr 2 and deviations from that mean for the genotypic classes with the higher protein alleles on chr 20 from PI468916, the *wp* allele from LN89-5320 or LN89-5322, or both higher protein alleles

Genetic background[a]	Locus[b]		*n*[c]	Seed yield (kg ha^{-1})	Seed protein (g kg^{-1})	Seed oil (g kg^{-1})	Maturity date[d]	Lodging[e] (1–5)	Plant height[f] (cm)
	chr 20[g]	wp[h]							
Loda	Low	*Wp*	18	3004	352	182	916	2.1	72
	Low	*wp*	17	−442	13**	−5	3**	−0.1	−5***
	High	*Wp*	17	−159	16***	−8***	−1	0.1	2*
	High	*wp*	13	−518**	31***	−14***	2	−0.1	−3**
Dwight	Low	*Wp*	18	3790	346	174	918	1.6	78
	Low	*wp*	22	−284***	4*	−2	−1	−0.1	−6***
	High	*Wp*	12	−279***	17***	−10***	−2**	0.2	3**
	High	*wp*	19	−603***	24***	−12***	−2***	0	−3***
LS93-0375	Low	*Wp*	17	4127	366	176	925	1.5	91
	Low	*wp*	19	−307*	0	0	2*	0	−5*
	High	*Wp*	16	−131	18***	−10***	−1	0	3**
	High	*wp*	23	−426**	25***	−13***	0	0	0
C1981	Low	*Wp*	12	4064	367	174	930	2.1	105
	Low	*wp*	11	−309***	0	0	1	0	−1
	High	*Wp*	11	−254***	25***	−13***	−2	0	8**
	High	*wp*	15	−567***	25***	−13***	−3*	−0.2	−3

Seed yield and protein and oil concentrations are reported on a 130 g kg^{-1} moisture basis
*, **, *** significant at the 0.05, 0.01, and 0.001 probability levels, respectively
a. Recurrent parent of population
b. Genotype of the genotypic class
c. Number of lines in the genotypic class
d. Characterized as the calendar date when 95% of pods have reached mature color (R8; Fehr et al. 1971) with September 1 equivalent to 901
e. Lodging is visually rated on a 1–5 scale with 1 = all plants erect and 5 = all plants prostrate
f. Distance between the soil line and the top node on the main stem
g. Genetic state at the chr 20 locus. 'low' is homozygous for the low protein allele, 'high' is homozygous for CHR20-PI
h. Genetic state at the *wp* locus. '*Wp*' is homozygous for the purple flower/low protein allele, and '*wp*' is homozygous for the pink flower/high protein allele

wp allele also had a variable effect in terms of significance when lines homozygous for *wp* were compared to lines with no high protein alleles on yield, maturity date, plant height within the Loda, Dwight, LS93-0375, and C1981 backgrounds. When lines contained both the *wp* allele and CHR20-PI, a significant increase in protein concentration was observed in all backgrounds in comparison to lines containing no high protein alleles; however, yield and oil concentration were significantly decreased.

Discussion

Although CHR20-PI has been studied for over two decades, detailed seed concentration and agronomic information on CHR20-D and the *wp* locus is more limited. Our study evaluated CHR20-PI, CHR20-D, and the *wp* alleles to determine whether they can be effectively used to improve seed composition in a breeding program targeted at improving the seed protein concentration in high-yield cultivar development.

For the most part, similar seed composition and yield trends were observed when the Danbaekkong high protein allele was introgressed into the Dwight and LD02-5025 backgrounds. This is not surprising because these two recurrent backgrounds not only have the same maturity but are also related with Dwight, a parent of LD02-5025. In both populations, lines containing the Danbaekkong high protein allele had decreased yield across and within environments and also had increased protein and decreased oil across and within all environments with the exception of Pontiac 2013 (Table 1). We do not have a good explanation for the inconsistent Pontiac 2013 results, but it may have to do with the growing environment at this location during 2013, as seed composition is influenced by numerous environmental conditions such as temperature and moisture (Dornbos and Mullen 1992; Gibson and Mullen 1996; Specht et al. 2001; Carrera et al. 2009). The influence of the growing environment is supported by Pontiac 2013 having the lowest average protein concentrations of the four environments where the population was grown.

In previous studies using elite germplasm from the southern USA, CHR20-D was shown to have an inconsistent effect on seed yield in southern environments (Harris 2001; Yates 2006). This contrasts with the results from both Danbaekkong populations in our study where the Danbaekkong high protein allele was consistently associated with significantly decreased yield. Furthermore, we observed a significant yield decrease even when there was no significant increase in protein concentration. A number of explanations for the apparent discrepancy between our study and the previous studies include, but are not limited to, environmental influence, genetic background, and genetic linkage.

While CHR20-D was consistently associated with decreased yields, CHR20-PI had a more variable effect on yield. Additionally, CHR20-PI significantly increased protein and decreased oil within all environments and populations. This consistency was not observed for CHR20-D where a significant effect on seed composition was not seen in the Pontiac 2013 environment for both populations. Because we did not introgress high-protein alleles of CHR20-D and CHR20-PI into the same genetic backgrounds and test them in the same environments, we cannot directly compare the effects of these two alleles. Therefore, we are unable to speculate on their allelic identity relationship based on this study. With that caveat noted, we did not observe that the CHR20-D allele had a numerically smaller effect on yield than did the CHR20-PI allele. A smaller effect may have been expected based on previous research with Danbaekkong in the southern USA (Harris 2001; Yates 2006).

For protein concentration, CHR20-PI was more consistent than the *wp* locus in increasing protein concentration across genetic backgrounds. In the C1981 population, lines containing the *wp* allele did not have a significant increase in protein concentration compared to lines containing no high protein alleles (Table 4). When the *wp* allele was stacked with CHR20-PI in this background, protein concentration was not numerically different than lines containing only CHR20-PI. Within the LS93-0375 background, the *wp* allele was ineffective in significantly increasing protein concentration on its own, but in combination with CHR20-PI, a significant increase in protein concentration in relation to lines with no high protein alleles was observed. Only in the Loda population were the *wp* allele and CHR20-PI numerically similar in their impact on protein concentration. Other than in the C1981 background, lines containing both the chr 20 and chr 2 protein-increasing alleles had on average the greatest protein concentration compared to lines in the other three possible genotypic groups. CHR20-PI increased protein concentration, but also decreased oil across genetic backgrounds and environments (Table 3). CHR20-PI also was associated with decreased yield and increased plant height variably across environments and genetic backgrounds (Tables 2, 3). Stacking *wp* in combination with CHR20-PI generally produced results that would be expected if two-locus interaction was not significant (i.e., the two alleles at each locus interacted in an additive fashion) for all traits across environments (Table 4). With the exception of the C1981 population, the combination of the high protein alleles at the chr 20 and chr 2 loci increased protein concentration to the greatest extent; however, this combination also decreased seed yield to the greatest extent within all genetic backgrounds including C1981. The reliability of CHR20-PI for increasing protein concentration would make it a better candidate than *wp* for a forward breeding application. However, if yield is the primary goal, neither allele would likely be a successful candidate in breeding program aimed at developing high-yield cultivars through a traditional marker-assisted selection (MAS) breeding scheme. Rapid improvements in genotyping and big data set analysis have led to recent protein and oil QTL mapping studies using diverse, large populations and with high density genetic markers (Hwang et al. 2014; Bandillo et al. 2015; Vaughn et al. 2014; Sonah et al. 2015; Phansak et al. 2016; Qi et al. 2016). While additional seed composition of QTL have been mapped in these studies, the chr 20 QTL region continues to be identified as having the largest effect on protein and oil concentration. Data from these studies can be used to better characterize and define the chr 20 QTL and ultimately clone it. As more information is generated about genes that control seed composition, this information can not only be used to dissect the genetic architecture of composition and generate more efficient markers for MAS, but also to provide insight into the relationship between seed composition and yield. Even with rapid advances in QTL mapping technologies and methods, QTL confirmation and evaluation studies remain important so that mapped QTL can be effectively incorporated into a breeding program to improve seed composition traits. Predictive modeling has shown promise to revolutionize plant breeding by improving genetic gain through a decrease of the length of breeding cycles and an increase in selection accuracy. Prediction accuracies over 0.60 have been reported for yield, protein, and oil, and it is assumed that these accuracies will further increase with improved statistical models and methods (Jarquin et al. 2014, 2016; Xavier et al. 2016). QTL mapping and evaluation studies can be important tools to aid breeders in selecting the most appropriate prediction model, making the model more robust, or assembling a strong training population. Overall, improved genomic selection techniques have potential to lead to the development of more high protein

and high-yield cultivars. The development of cultivars with improved yield and protein concentration continues to be challenging due to the negative relationship between the two traits. The QTL evaluated in this study, and in other studies where protein and yield were both evaluated, provide genetic evidence for this negative correlation (Hegstad et al. 2000; Sebolt et al. 2000; Chung et al. 2003; Nichols et al. 2006). We cannot demonstrably document whether the impact on both protein and oil of the two alleles at the chr 20 and chr 2 QTLs that we studied arose from single-locus pleiotropy or two-locus linkage. However, the inability of researchers to separate the effect of the QTL on both traits and the high energy cost of producing protein suggests that it is likely pleiotropy. The continued evaluation of QTL combined with advancements in genetic technologies could help us better understand the genetic relationships among seed components and lead to better strategies to develop cultivars with increased protein concentration and yield.

Author contributions — LFB conducted genetic and field experiments, analyzed data, and drafted manuscript; TIP conducted genetic and field experiments, analyzed data, interpreted results, and edited manuscript; JES conducted field experiments and edited manuscript; BWD designed and organized project and edited manuscript.

Acknowledgments — This research was supported by funding from the United Soybean Board (USB) to BWD and LFB.

Conflict of interest — The authors declared no conflicts of interest.

References

Abney TS, Crochet WD (2003) Uniform soybean tests northern states, 2003. USDA-ARS, West Lafayette

Abney TS, Crochet WD (2006) Uniform soybean tests northern states, 2006. USDA-ARS, West Lafayette

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A (2015) A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8(3). doi: 10.3835/plantgenome2015.04.0024

Bell-Johnson B, Garvey G, Johnson J, Lightfoot D, Meksem K (1998) Biotechnology approaches to improving resistance to SCN and SDS: methods for high throughput marker assisted selection. Soybean Genet Newsl 25:115–117

Bolon YT, Joseph B, Cannon SB, Graham MA, Diers BW, Farmer AD, May GD, Muehlbauer GJ, Specht JE, Tu ZJ, Weeks N (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol 10:41

Brim CA, Burton JW (1979) Recurrent selection in soybeans II: selection for increased percent protein in seeds. Crop Sci 19:494–498

Brummer E, Graef G, Orf J, Wilcox J, Shoemaker R (1997) Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci 37:370–378

Carrera C, Martínez MJ, Dardanelli J, Balzarini M (2009) Water deficit effect on the relationship between temperature during the seed fill period and soybean seed oil and protein concentrations. Crop Sci 49:990–998

Cary TR, Diers BW (2007) Northern regional soybean cyst nematode tests. University of Illinois, Urbana

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

Cober ER, Voldeng HD (2000) Developing high-protein, high-yield soybean populations and line. Crop Sci 40:39–42

Cregan P, Quigley C (1997) Simple sequence repeat DNA marker analysis. In: Caetano-Anolles G, Gresshoff PM (eds) DNA markers: protocols, applications, and overviews. Wiley, New York, pp 173–185

Crochet WD, Hughes T (2011) Uniform soybean tests northern states, 2009. USDA-ARS, West Lafayette

Cromwell GL (2012) Soybean meal: an exceptional protein source. Soybean Meal InfoCenter. http://www.soymeal.org/ReviewPapers/SBMExceptionalProteinSource.pdf Accessed 30 Jan 2017

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

Diers BW, Cary TR, Thomas DJ, Nickell CD (2006) Registration of 'LD00-3309' soybean. Crop Sci 46:1384

Dornbos DL, Mullen RE (1992) Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. J Am Oil Chem Soc 69:228–231

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

Fasoula VA, Harris DK, Boerma HR (2004) Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci 44:1218–1225

Fehr W, Caviness C, Burmood D, Pennington J (1971) Stage of development descriptions for soybeans, *Glycine Max* (L.) Merrill. Crop Sci 11:929–931

Gibson LR, Mullen RE (1996) Soybean seed composition under high day and night growth temperatures. J Am Oil Chem Soc 73:733–737

Hanson WD, Leffel RC, Howell RW (1961) Genetic analysis of energy production in the Soybean. Crop Sci 1:121–126

Harris DK (2001) Genetic improvement of soybean seed traits and resistance to bud blight and root-knot nematodes. MS thesis, University of Georgia

Hartwig EE, Kilen TC (1991) Yield and composition of soybean seed from parents with different protein, similar yield. Crop Sci 31:290–292

Hegstad JM, Vodkin LO, Nickell CD (2000) Genetic and agronomic evaluation of *wp-m* in soybean. Crop Sci 40:346–351

Holbrook CC, Burton JW, Carter TE (1989) Evaluation of recurrent restricted index selection for increasing yield while holding seed protein constant in soybean. Crop Sci 29:324–329

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

Jarquin D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genom 15:740

Jarquin D, Specht J, Lorenz A (2016) Prospects of genomic prediction in the USDA soybean germplasm collection: historical data

creates robust models for enhancing selection of accessions. G3 Genes Genomes Genet 6:2329–2341
Keim P, Olson R, Shoemaker R (1988) A rapid protocol for isolating soybean DNA. Soybean Genet Newsl 12:150–152
Kerley MS, Allee GL (2003) Modifications in soybean seed composition to enhance animal feed use and value: moving from a dietary ingredient to a functional dietary component. AgBioForum 6:14–17
Kim SD, Hong EH, Kim YH, Lee SH, Seong YK, Park KY, Lee YH, Hwang EH, Park EH, Kim HS, Ryu YH, Park RK, Kim YS (1996) A new high protein and good seed quality soybean variety "Danbaegkong" (in Korean). RDA J Agric Sci Upland Ind Crops 38:228–232
Kim M, Schultz S, Nelson R, Diers BW (2015) Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci 56:219–225
Lee SH, Bailey MA, Mian RM, Carter TE, Shipe ER, Ashley DA, Parrot WA, Hussey RS, Boerma HR (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor Appl Genet 93:649–657
Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006) SAS® for mixed models, 2nd edn. SAS Institute Inc, Cary
Liu K (1997) Soybeans: chemistry, technology, and utilization. Chapman and Hall, New York
Mian R, McHale L, Li Z, Dorrance AE (2017) Registration of 'Highpro1' Soybean with high protein and high yield developed from a north × south cross. J Plant Reg 11:51–54
Nichols D, Glover K, Carlson S, Specht J, Diers B (2006) Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci 46:834–839
Nickell CD, Thomas DJ, Cary TR, Heavner D (1996) Registration of 'Macon' soybean. Crop Sci 36:1410
Nickell CD, Noel GR, Cary TR, Thomas DJ, Leitz RA (1998) Registration of 'Dwight' soybean. Crop Sci 38:1398
Nickell CD, Noel GR, Cary TR, Thomas DJ, Diers BW (2001) Registration of 'Loda' soybean. Crop Sci 41:589–590
Nowling GL (2001) Uniform soybean tests northern states, 2001. USDAARS, West Lafayette
Phansak P, Soonsuwon W, Hyten DL, Song Q, Cregan PB, Graef GL, Specht JE (2016) Multi-population selective genotyping to identify soybean (*Glycine max* (L.) Merr.) seed protein and oil QTLs. G3 Genes Genomes Genet 6:1635–1648
Qi Z, Pan J, Han X, Qi H, Xin D, Li W, Mao X, Wang Z, Jiang H, Liu C, Hu Z, Hu G, Zhu R, Chen Q (2016) Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. Mol Breed 36:1–16
Recker JR, Burton JW, Cardinal A, Miranda L (2014) Genetic and phenotypic correlations of quantitative traits in two long-term randomly mated soybean populations. Crop Sci 54:939–943
SAS Institute (2016) The SAS system for Microsoft Windows, release 9.4. SAS Inst, Cary, NC
Schmidt ME, Klein JH (2002) Registration of 'LS93-0375' soybean. Crop Sci 42:1375
Sebolt A, Shoemaker R, Diers B (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40:1438–1444
Shimura E, Hanson WD (1970) Covariance analysis involving energy production and distribution among seed fractions by soybean genotypes, *Glycine max* (L.) Merrill. Crop Sci 10:242–246
Sinclair TR, de Wit CT (1975) Photosynthate and nitrogen requirements for seed production by various crops. Science 189:565–567
Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2015) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J 13:211–221
SoyBase and the Soybean Breeder's Toolbox (2017) Soybean breeder's toolbox genetic map information. http://www.soybase.org Accessed 30 Jan 2017
Specht J, Chase K, Macrander M, Graef G, Chung J, Markwell J, Germann M, Orf J, Lark K (2001) Soybean response to water. Crop Sci 41:493–509
Stephens PA, Nickell CD (1992) Inheritance of pink flower in soybean. Crop Sci 32:1131–1132
Stephens PA, Nickell CD, Vodkin LO (1993) Pink flower color associated with increased protein and seed size in soybean. Crop Sci 33:1135–1137
USDA-ARS (2017) National Genetic Resources Program. Germplasm Resources Information Network-(GRIN). National Germplasm Resources Laboratory, Beltsville. https://npgsweb.ars-grin.gov/gringlobal Accessed 12 Jan 2017
Vaughn JN, Nelson RL, Song Q, Cregan PB, Li Z (2014) The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3 Genes Genomes Genet 4:2283–2294
Wang D, Shi J, Carlson S, Cregan P, Ward R, Diers B (2003) A low-cost, high-throughput polyacrylamide gel electrophoresis system for genotyping with microsatellite DNA markers. Crop Sci 43:1828–1832
Wang X, Jiang GL, Green M, Scott RA, Son Q, Hyten DL, Cregan PB (2014) Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. Mol Genet Genom 289:935–949
Warrington CV, Abdel-Haleem H, Hyten DL, Cregan PB, Orf JH, Killam AS, Bajjalieh N, Li Z, Boerma HR (2015) QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor Appl Genet 128:839–850
Wehrmann VK, Fehr WR, Cianzio SR, Cavins JF (1987) Transfer of high seed protein to high-yielding soybean cultivars. Crop Sci 27:927–931
Wilcox JR, Cavins JF (1995) Backcrossing high seed protein to a soybean cultivar. Crop Sci 35:1036–1041
Wilcox JR, Shibles RM (2001) Interrelationships among seed quality attributes in soybean. Crop Sci 41:11–14
Wilson RF (2004) Seed composition. In: Boerma HR, Specht JE (eds) Soybeans: improvement, production, and uses, 3rd edn. Agron Monogr No 16. ASA, CSSA, and SSSA, Madison, pp 621–677
Wilson RF (2008) Soybean: Market driven research needs. In: Stacey G (ed) Genetics and genomics of soybean, 2nd edn. Springer, New York, pp 3–14
Xavier A, Muir WM, Rainey KM (2016) Assessing predictive properties of genome-wide selection in soybeans. G3 Genes Genomes Genet 6:2611–2616
Yates JL (2006) Use of diverse germplasm to improve peanut rootknot nematode resistance and seed protein content in soybean. Dissertation, University of Georgia
Zabala G, Vodkin LO (2005) The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell 17:2619–2632