4-1-2022

Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20

Christina E. Fliege
University of Illinois Urbana-Champaign

Russell A. Ward
University of Illinois Urbana-Champaign

Pamela Vogel
University of Nebraska–Lincoln

Hanh Nguyen
University of Nebraska–Lincoln, hnguyen8@unl.edu

Truyen Quach
University of Nebraska–Lincoln, tquach2@unl.edu

See next page for additional authors

the plant journal

The Plant Journal (2022) 110, 114–128 doi: 10.1111/tpj.15658

Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20

Christina E. Fliege[1], Russell A. Ward[1,†], Pamela Vogel[2,‡], Hanh Nguyen[3], Truyen Quach[3], Ming Guo[2], João Paulo Gomes Viana[1], Lucas Borges dos Santos[1], James E. Specht[2], Tom E. Clemente[2], Matthew E. Hudson[1] and Brian W. Diers[1]*

[1]Department of Crop Sciences, University of Illinois, 1101 W. Peabody Dr., Urbana, IL 61801, USA,
[2]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA, and
[3]Center for Plant Science Innovation, University of Nebrasaka-Lincoln, Lincoln, NE 68583, USA

Received 16 July 2021; accepted 28 December 2021; published online 3 January 2022.
*For correspondence (e-mail bdiers@illinois.edu).
†Present address: Syngenta Seeds Inc., Aurora, SD 57002, USA
‡Present address: Pairwise Company, Durham, NC 27701,USA

SUMMARY

Soybean [*Glycine max* (L.) Merr.] is a unique crop species because it has high levels of both protein and oil in its seed. Of the many quantitative trait loci (QTL) controlling soybean seed protein content, alleles of the cqSeed protein-003 QTL on chromosome 20 exert the greatest additive effect. The high-protein allele exists in both cultivated and wild soybean (*Glycine soja* Siebold & Zucc.) germplasm. Our objective was to fine map this QTL to enable positional-based cloning of its underlying causative gene(s). Fine mapping was achieved by developing and testing a series of populations in which the chromosomal region surrounding the segregating high- versus low-protein alleles was gradually narrowed, using marker-based detection of recombinant events. The resultant 77.8 kb interval was directly sequenced from a *G. soja* source and compared with the reference genome to identify structural and sequence polymorphisms. An insertion/deletion variant detected in *Glyma.20G85100* was found to have near-perfect +/− concordance with high/low-protein allele genotypes inferred for this QTL in parents of published mapping populations. The indel structure was concordant with an evolutionarily recent insertion of a TIR transposon into the gene in the low-protein lineage. Seed protein was significantly greater in soybean expressing an RNAi hairpin downregulation element in two independent events relative to control null segregant lineages. We conclude that a transposon insertion within the CCT domain protein encoded by the *Glyma.20G85100* gene accounts for the high/low seed protein alleles of the cqSeed protein-003 QTL.

Keywords: soybean, *Glycine max* (L.) Merr., seed protein, QTL, gene cloning, fine mapping.

INTRODUCTION

Soybean is an important source of protein and oil for both animal and human consumption. It has the highest seed protein concentration (averaging 400 g kg^{-1}, 40% on a dry weight basis) when compared with other major legume crop species (average range of 200–300 g kg^{-1} seed protein), which in turn surpasses cereal crop species (average range of 80–150 g kg^{-1}) (Liu, 1997). Most soybean protein in the market is a co-product of soybean oil extraction. This soybean meal is a highly digestible feedstuff that provides the essential amino acids needed for animal growth (Cromwell, 2012). Despite being used as a source of oil and protein, the soybean market value is driven primarily by the soybean meal fraction rather than the oil fraction. In 2018, 70% of the world's protein meal consumption came from soybean, totaling to 235.4 million metric tons (American Soybean Association, 2019).

Seed protein concentration in soybean is quantitatively inherited (Burton, 1985; Wilcox, 1985). Many seed protein-controlling quantitative trait loci (QTL) have now been identified (Grant et al., 2010). The Soybase website (Grant et al., 2010) lists 248 marker associations with seed protein concentration that since the early 1990s to date (June 2021) were detected in biparental populations. These QTL have been loosely mapped to multiple overlapping regions on each of the 20 soybean chromosomes, but the majority are likely redundant detections of many fewer genuine QTL. For this reason, breeders have used the acronym "cq" (confirmed QTL) to highlight those QTL whose map position has been experimentally confirmed and approved

by the Soybean Genetics Committee. A seed protein QTL on soybean chromosome 20 (formerly linkage group I), now known as cqSeed Protein-003, has been one of the most widely studied protein QTL due to its large additive effect on seed protein. When this QTL is mapped in a population, QTL for seed oil (cqSeed Oil-004), seed yield (cqSeed Yield-001), and seed mass (cqSeed weight-003) are frequently identified in the same genetic region likely because of pleiotropy (Nichols et al., 2006). Plants homozygous for the high-protein QTL allele have been shown to have protein concentration increases >20 g kg^{-1} and decreases in oil of approximately 10 g kg^{-1} compared with plants with the alternative allele (Diers et al. 1992; Brummer et al. 1997; Sebolt et al. 2000; Csanádi et al. 2001). To increase our understanding of the underlying genetic control of cqSeed protein-003, fine mapping followed by candidate gene identification and cloning is needed (Salvi and Tuberosa 2007).

Both cqSeed protein-003 and cqSeed oil-004 were first mapped with RFLP markers by Diers et al. (1992). This QTL was discovered in a population derived from crossing the *Glycine max* line A81-356022 with the *Glycine soja* plant introduction PI 468916. Based on the marker analysis, the allele from *G. soja* was associated with a protein increase of 24 g kg^{-1} (Diers et al., 1992). After this QTL was identified, the high-protein allele from *G. soja* was introgressed into different genetic backgrounds, which confirmed that the allele could be used to increase the seed protein content, but it was also found to be associated with lower yield (Sebolt et al., 2000). Fine mapping of cqSeed protein-003 was initiated by Nichols et al. (2006) in a study that narrowed the QTL map position to a 3 cM region. Bolon et al. (2010) then used a large set of simple sequence repeat (SSR) markers to narrow the QTL interval to an 8.4-Mbp region between the markers Sat_174 and ssrpqtl_38.

To date, cqSeed protein-003 and cqSeed oil-004 QTLs have been mapped in multiple biparental crosses (Brummer et al., 1997; Chung et al., 2003; Kim et al., 2016; Lu et al., 2012; Phansak et al., 2016; Reinprecht et al., 2006; Tajuddin et al., 2003; Wang et al., 2014; Warrington et al., 2015). Genome-wide association mapping studies (GWAS) were also used to narrow the base pair interval the QTL maps. Hwang et al. (2014) conducted a GWAS using 42 368 single nucleotide polymorphisms (SNPs) in a genetically diverse set of 298 lines. They narrowed the candidate gene region to a 2.4-Mbp interval located at 28.7–31.1 Mbp (Gmax2.0 assembly). In a GWAS conducted by Vaughn et al. (2014), the candidate gene region was positioned in the approximately 1-Mbp region between 32.1 and 33.1 Mbp (Gmax2.0 assembly) using mostly maturity group (MG) V accessions from South Korea. Bandillo et al. (2015) used GWAS to analyze 12 000 accessions (all MGs) from the USDA Soybean Germplasm Collection using 36 513 SNPs, and provided evidence that the candidate gene resided in the approximately 2.4-Mbp interval between 30.7 and 33.1 Mbp (Gmax2.0 assembly).

To detect soybean QTL with large effects on seed protein, Phansak et al. (2016) used a multiple-population selective genotyping strategy by mating 48 accessions (ranging from MG 000 to IV) with high yielding cultivars of the same MG and evaluating $F_{2:3}$ progeny. The accessions all had high-protein content (412–458 g kg^{-1} on a 13% moisture basis) and the cultivars had ordinary protein concentrations (332–374 g kg^{-1}). After sorting the $F_{2:3}$ progeny for seed protein concentration, just the upper and lower deciles were genotyped with SNP markers. A protein QTL was detected at the same position as cqSeed protein-003 in 27 of the 48 matings, indicating that in the high-protein germplasm, the high-protein allele of this QTL predominates.

In our current study, further fine mapping of cqSeed protein-003 was conducted with near isogenic lines derived from backcrossing the *G. soja* protein allele into the A81-356022 background. The narrowed candidate region was sequenced and assembled using Illumina technology. Polymorphisms in candidate genes from the interval were tested against a panel of soybean genotypes known either to have or not to have the high-protein allele in the cqSeed protein-003 interval. Ultimately, an insertion/deletion polymorphism in the CCT-domain protein-encoding gene *Glyma.20G85100* was identified as the most likely candidate. The role of this gene in controlling seed protein content was subsequently confirmed by stable transformation of soybean lines.

RESULTS

Fine mapping the QTL interval

The first round of mapping was done using a set of BC_5F_7 populations developed from BC_5F_6 plants that were selected for recombination between markers Satt239 and ssrpqtl_18 (Table 1). Previous research in our laboratory indicated that these two markers bracketed cq-Seed protein-003 (Sebolt et al. 2000; Nichols et al. 2006). BC_5F_7 plants in these populations were tested with a segregating marker and field evaluated for seed protein content in 2008 followed by the evaluation of $BC_5F_{7:8}$ lines in 2009. The protein and marker results were analyzed to test for a statistical association between the protein and marker data to determine whether the QTL was in the segregating or non-segregating interval in each population. For example, a significant association was found in BC_5F_7 population 4, which showed that the QTL was the segregating interval below ssrpqtl_17 (Table 1). Conversely, no association was found in population 2, which indicated that the QTL is in the non-segregating interval below ssrpqtl_17. By combining the results across the 13 populations, we concluded that the QTL was within a 5.5-Mbp interval between

Table 1 List of 19 soybean chromosome 20 markers (Bolon et al., 2010), ordered by their base pair (bp) positions (Williams 82 assemblies – see www.soybase.org), that were used to characterize 13 populations of BC_5F_7 plants and their descendent $BC_5F_{7:8}$ lines in the first round of fine mapping (Figure S1). Each of the 13 BC_5F_7 populations was developed from a separate BC_5F_6 plant and the genotype of each of these parent plants for the chromosome 20 markers are show below. If the BC_5F_6 plant thepopulation was developed from was homozygous for the donor parent (*G. soja*) high protein allele, genotypic code B was used, A was used when the plant was homozygous for the recurrent parent (*G. max*) allele, and H when the plant was heterozygous. Probabilities are given for whether the marker segregation in each population was significantly associated with protein based on field tests and an arrow is pointed in the direction, relative to crossovers, to denote where the QTL was located. The region where these tests show the QTL is located is shaded in grey

			BC_5F_7 Population No.												
Marker name	GMax1.01 bp position	GMax2.0 bp position	2	3	10	14	22	4	6	13	17	20	11	15	19
Satt239	24,129,682	25,275,083	H	H	H	H	H	A	A	B	A	A	B	B	A ↓
ssrpqtl_4	24,812,334	25,971,714	H	H	H	H	H	A	A	B	A	A	B	B	H
ssrpqtl_8	25,751,901	26,920,157	H	H	H	H	H	A	A	B	A	A	B	B	H
ssrpqtl_11	26,270,814	27,439,056	H	H	H	H	H	A	A	B	A	A	B	B	H
ssrpqtl_13	26,444,803	27,606,228	H	H	H	H	H	A	A	B	A	A	B	B ↓	H
ssrpqtl_14	26,538,403	27,699,841	H	H	H	H	H	A	A	B	A	A	B	H	H
ssrpqtl_15	26,542,454	27,703,952	H	H	H	H	H	A	A	B	A	A	B ↓	H	H
ssrpqtl_16	26,609,299	27,770,740	H	H	H	H	H	A	A	B	A	A	H	H	H
ssrpqtl_17	26,649,308	27,810,743	H ↓	H ↓	H ↓	H ↓	H ↓	A ↓	A ↓	B ↓	B ↓	A ↓	H	H	H
ssrpqtl_18	26,958,336	28,124,804	B	A	B	A	B	H	H	H	H	H	H	H	H
ssrpqtl_25	30,489,918	31,627,304	B	A	B	A	B	H	H	H	H	H	H	H	H
ssrpqtl_29	31,787,239	32,934,791	B	A	B	A	B	H	H	H	H	H	H	H	H
ssrpqtl_32	31,992,972	33,141,346	B	A	B	A	B	H	H	H	H	H	H	H	H
ssrpqtl_33	32,022,042	33,170,565	B	A	B	A	B	H ↑	H ↑	H ↑	H ↑	H	H ↑	H ↑	H
ssrpqtl_34	32,178,223	33,326,612	A	A	A	A	B	A	A	A	A	H	A	A	H
ssrpqtl_35	32,216,450	33,359,151	A	A	A	A	B	A	A	A	A	H	A	A	H
ssrpqtl_36	32,384,780	33,526,075	A	A	A	A	B	A	A	A	A	H	A	A	H
ssrpqtl_37	32,717,564	33,858,592	A	A	A	A	B	A	A	A	A	H ↑	A	A	H ↑
ssrpqtl_38	32,910,185	34,049,358	A	A	A	A	A	A	A	A	A	A	A	A	A
2008 Prob > F[†]			NS	NS	NS	NS	NS	*	**	*	**	*	**	**	**
2009 Prob > F			NS	NS	NS	NS	NS	**	**	-	-	**	-	-	-

[†] Significance of the marker association test in each population based on field testing of each population of BC_5F_7 plants in the field 2008 and each population of $BC_5F_{7:8}$ lines in the field in 2009. NS not significant,
*Significant at 0.05 probability,
**Significant at 0.01 probability, and dash (–) not tested.

ssrqtl_17 and ssrqtl_34 on chromosome 20 (Gmax2.0) (Table 1, Table S1). This interval was novel in that was located on the distal side of where we originally hypothesized the QTL was positioned based on our preliminary data when we started the study.

A second round of population development and fine mapping was commenced to narrow the QTL interval further (Figure S1). Eleven populations were developed from BC_5F_8 plants that had recombination events across the QTL interval and subsets of these populations were grown in the field as BC_5F_9 plants in 2011 and as $BC_5F_{9:10}$ lines in 2012 and $BC_5F_{9:11}$ lines 2013 (Table 2). As in the first round, each plant in the populations was genotyped with a segregating marker and the seed harvested from the plants and lines from the populations were analyzed for protein content to determine which population was segregating for the QTL (Table S2). The results across the populations, with the exception of one test, placed cqSeed protein-003 on chromosome 20 between BARCSOYSSR_20_0670 and BARCSOYSSR_20_0674 (Table 2). This corresponded to a 77.8-kb region between 31 744 150 and 31 821 947 bp based on the Wm82.a2.v1 (Gmax2.0) assembly (SoyBase, https://soybase.org).

The only inconsistent results were from BC_5F_9 Population 7 evaluated in 2013 (Table 2 and Table S2). The significant protein-marker association in this test indicated that the QTL was above the marker 20_0657, which was inconsistent with the other 2 years it was tested and other populations evaluated in this round. To settle this inconsistency, seven subpopulations were developed from

Table 2 List of 17 soybean chromosome 20 markers, ordered by their bp positions (Williams 82 assemblies – see www.soybase.org), that were used to characterize 11 populations of BC_5F_9 plants grown in 2011 and their descendent $BC_5F_{9:10}$ and $BC_5F_{9:11}$ lines grown in 2012 and 2013, respectively, in the second round of fine mapping. Each of the 11 BC_5F_9 populations was developed from a separate BC_5F_8 plant and the genotype of each of these parent plants for the chromosome 20 markers are show below. If the BC_5F_9 plant the population was developed from was homozygous for the donor parent (*G. soja*) high protein allele, genotypic code B was used, A was used when the plant was homozygous for the recurrent parent (*G. max*) allele, and H when the plant was heterozygous. Probabilities are given for whether the marker segregation in each population was significantly associated with protein based on field tests and an arrow is pointed in the direction, relative to crossovers, that the QTL is located. The region where these tests show the QTL is located is shaded in grey. The markers are from Bolon et al. (2010) and Song et al. (2010) and for brevity, the prefix BARCSOYSSR was dropped from the names of markers from Song et al.

			BC_5F_9 Population No.										
Marker name	Gmax1.01 bp Position	Gmax2.0 bp Position	3	5	7	9	10	15	26	27	28	31	34
ssrpqtl_17	26,649,365	Not avail.	H	H	H	H	H	H	A	A	B	A	A ↓
20_0599	26,829,294	27,991,409	H	H	H	H	H	H	A	A	B	A	H
20_0616	27,811,875	28,974,676	H	H	H	H	H	H	A	A	B	A ↓	H
20_0617	27,877,620	29,040,539	H	H	H	H	H	H	A	A	B	H	H
20_0636	28,972,334	30,134,877	H ↓	H	H	H	H	H	A	A	B	H	H
20_0647	29,643,301	30,793,572	A	H	H	H	H	H	A	A	B	H	H
20_0650	29,758,405	30,909,346	A	H ↓	H	H	H	H	A	A	B	H	H
20_0655	30,052,089	31,198,164	A	B	H ↓	H	H	H	A	A	B	H	H
20_0657	30,187,698	31,333,773	A	B	A	H	H	H	A	A	B	H	H
20_0667	30,489,863	31,627,414	A	B	A	H	H	H	A	A	B ↓	H	H
20_0668	30,517,621	31,655,159	A	B	A	H	H	H	A	A	H	H	H
20_0670	30,606,609	31,744,150	A	B	A	H ↓	H	H	A	A ↑	H	H	H
20_0674	30,684,404	31,821,947	A	B	A	B	H ↑	H	A ↑	H	H	H	H
20_0678	30,754,018	31,891,560	A	B	A	B	A	H	H	H	H	H	H
20_0715	32,030,122	33,178,717	A	B	A	B	A	H ↑	H	H	H	H	H
20_0718	32,178,222	33,326,648	A	B	A	B	A	A	H	H	H	H	H
ssrpqtl_34	32,178,303	Not avail.	A	B	A	B	A	A	H	H	H	H	H
2011 Prob > F[†]			NS	NS	NS	-	**	**	-	-	**	**	**
2012 Prob > F					NS	NS	**		NS	NS	**	-	-
2013 Prob > F					**	NS	-	-	-	NS	**	-	-

[†]Significance of the marker association test based on field testing of each population of BC_5F_9 plants in the field 2011 and each population of $BC_5F_{9:10}$ and $BC_5F_{9:11}$ lines in the field in 2012 and 2013, respectively. NS denotes not significant,
*Significant at 0.05 probability,
**Significant at 0.01 probability, and dash (–) not tested.

Population 7 that had the same recombination breakpoint as the original plant used to develop Prol-7 and segregate for the same interval. No significant marker association was found in tests conducted in 2015 of any of the seven subpopulations (Table S3), thus refuting the initial inconsistency and firmly supporting the finding that the QTL is below BARCSOYSSR_20_0670.

Identification of candidate polymorphisms in interval

Three candidate genes were identified within the 77.8-kb region based on the Gmax2.0 map assembly (SoyBase, https://soybase.org). These genes were: *Glyma.20g085000*, encoding a component of the oligomeric Golgi complex; *Glyma.20g085100*, a CCT motif protein encoding gene; and *Glyma.20g085200*, encoding a zinc-ion binding LIM domain protein. To analyze this region in depth, we took advantage of the whole genome *de novo* assembly of PI 468916 DNA previously described by Butler et al. (2021). The assembly within the interval covered all annotated genes within large contigs. Along with additional reads and polymerase chain reaction (PCR) products, the assembly was analyzed and compared with the Williams 82 reference genome sequence by alignment and visual inspection to identify structural as well as sequence polymorphisms. When the

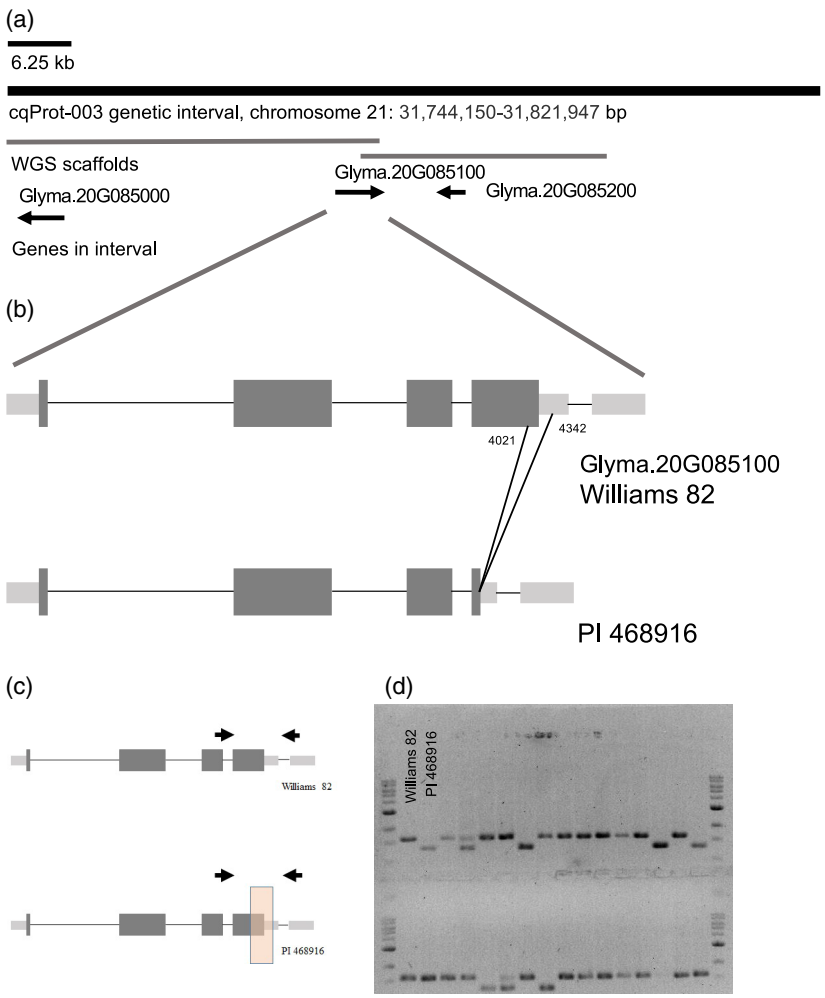

Figure 1. Molecular characterization of the genetic interval of the cqSeed protein-003 interval.
(a) Diagram showing the position of assembled scaffolds within the interval, and annotated genes. Region to the right of the interval contains a repetitive sequence that does not code for protein.
(b) Expanded view of *Glyma.20G085100*, showing the respective 321 bp insertion in Williams 82 compared with the PI 468916 sequence.
(c) Diagram showing the primer combination used to develop a codominant marker for this polymorphism.
(d) Agarose gel showing genotyping of part of the population of soybean accessions in Table 3 using the marker.

aligned genomic sequence of PI 468916 and Williams 82 were compared for the three candidate genes, large insertion–deletion mutations (indels) were identified in *Glyma.20g085100* and *Glyma.20g085200*, but only an exonic synonymous mutation was found in *Glyma.20g085000*.

In *Glyma.20g085100* (Figure 1a) the comparison revealed a 321-bp indel that removed an annotated exon from PI 468916, and for *Glyma.20g085200*, there was an 8-kb indel that resulted in the removal of almost the entire annotated gene in PI 468916 (Figure 1b). PCR-based size markers were developed for both indels, and these markers were used to screen a panel of 53 parents with known alleles for cqSeed protein-003 based on previous QTL mapping results (Table 3). For the 8 kb deletion in *Glyma.20g085200*, there was no association between it and the presence or absence of the high-protein allele in the parent panel indicating that *Glyma.20g085200* was likely not the causal gene underlying cqSeed protein-003 (Table 3).

In contrast, an association was found between the *Glyma.20g085100* indel and the presence or absence of the high-protein allele in panel. Of the 32 parents that had been identified as having a high-protein allele at cqSeed protein-003 based on previous mapping studies (Table 3), all had the shorter version of the PCR fragment (PI 468916 allele) at *Glyma.20g085100*. Of the 21 parents that had been reported to have a low-protein version of cqSeed protein-003 in mapping studies (Table 3), 18 had the longer PCR fragment (Williams 82 type sequence), whereas three had the shorter fragment (Figure 1c,d). The three exceptions (PI 407877B, PI 423954, and PI 424148; Table 3) were among the high-protein parents mated to a low-protein parent in a multi-parent selective genotype QTL mapping study conducted by Phansak et al. (2016). Note that these lines were not explicitly shown to have the low-protein genotype, but the authors did not detect a statistically significant segregation of a chromosome 20 protein QTL in those three exceptions.

Using 50-K SNP data from the region, along with PCR genotyping at the marker in a sample set of diverse accessions (Tables 3 and S4), we performed haplotype, linkage disequilibrium (LD), and principal components analysis (PCA). When the *Glyma.20g085100* marker was compared with four markers flanking the gene for the 238 haplotyped

Table 3 List of soybean germplasm accessions that have been reported to be homozygous for either the high (H) or low (L) protein allele for a QTL in the chromosome 20 cqSeed protein-003 QTL region. This characterization was based on the use of each accession as a parent in a biparental population in a QTL mapping study (for details, see the cited report in the Source column). A polymerase chain reaction-based marker diagnostic assay was used to characterize each listed accession as to whether it possessed the PI 468916 *Glycine soja* (+) allele or the Williams 82 *Glycine max* (−) allele at each the two *Glyma. 20* genes that were inferred in the present study to be potential candidates for the QTL. Haplotypes were determined using the marker from *Glyma.20G085100* and the two closest markers flanking each side of the gene. The haplotype designations are defined in the Experimental procedures section

Accession name	Accession number	Source[a]	Allele at cqSeed protein-003	*Glyma. 20g085100*	*Glyma. 20g085200*	Haplotype
–	PI 468916	Diers et al. (1992)	H	+	+	5
–	PI 326582A	Chaky et al. (2003)	H	+	+	6
Kosodiguri Extra Early	FC 30687	P5	H	+	+	5
Akazu	PI 91725–4	P34	H	+	+	5
N-34	PI 153293	P6	H	+	+	5
V-4	PI 153296	P1	H	+	+	5
V-6	PI 153297	P14	H	+	–	5
V-14	PI 153301	P12	H	+	–	5
V-16	PI 153302	P9	H	+	+	5
No. 51	PI 154196	P23	H	+	+	5
–	PI 159764	P10	H	+	+	5
No. 58	PI 181571	P20	H	+	+	2
Bitterhof	PI 189880	P13	L	–	+	1
Geant Vert	PI 189963	P2	H	+	–	5
Kariho-takiya	PI 243532	P36	H	+	+	2
No. 17	PI 253666A	P40	H	+	+	2
Wasedaizu No. 1	PI 261469	P19	H	+	–	2
–	PI 340011	P35	H	+	–	2
Oshimashirome	PI 360843	P39	L	–	–	1
Ronset 4	PI 372423	P4	H	+	–	5
KAERI-GNT 310–1	PI 398516	P33	L	–	–	1
KAS 330–9-1	PI 398704	P44	H	+	–	2
–	PI 398881	Diers et al. (in prep)	L	–	–	1
KLS 630–1	PI 398970	P45	L	–	–	2
Huaj an si er dian	PI 404188A	Diers et al. (in prep)	L	–	–	1
KAS 330–9-2	PI 407773B	P47	H	+	+	2
ORD 8113	PI 407788A	P41	H	+	–	2
–	PI 407823	P46	L	–	–	1
KAERI 511–11	PI 407877B	P43	(L)	(+)	–	2
KAS 640–7	PI 408138 C	P37	H	+	–	
Saikai 1	PI 423942	P28	H	+	–	2
Saikai 18	PI 423948A	P29	H	+	–	2
Saikai 20	PI 423949	P25	H	+	–	2
Shirome	PI 423954	P22	(L)	(+)	–	2
Shirome	PI 424148	P21	(L)	(+)	–	2
KAS 239–4	PI 424286	P42	L	–	–	2
Backchung No. 42	PI 427136	Diers et al. (in prep)	L	–	–	1
Choseng No. 1	PI 427138	P18	H	+	+	2
Seuhae No. 20	PI 427141	P26	H	+	+	2
DV-147	PI 437088A	P24	H	+	–	2
VIR 249	PI 437112A	P30	L	–	–	1
VNIISC-4	PI 437169B	Diers et al. (in prep)	L	–	–	1
Sjuj-dja-pyn-da-do	PI 437716A	P27	H	+	–	2
Ronest 4	PI 438415	P11	H	+	+	5

(continued)

Table 3 (continued)

Accession name	Accession number	Source[a]	Allele at cqSeed protein-003	*Glyma. 20g085100*	*Glyma. 20g085200*	Haplotype
Szu yueh pa	PI 445845	P32	H	+	–	2
KAS 578–1	PI 458256	P48	L	–	+	1
NS-20	PI 518751	Diers et al. (in prep)	L	–	–	1
Provar	PI 548608	P31	L	–	–	1
Fen dou 14	PI 561370	Diers et al. (in prep)	L	–	–	1
Williams 82	PI 518671	Kim et al. (2016)	L	–	–	1
A81-355012	A81-355012	Diers et al. (1992)	L	–	–	
Ina	PI 606749	Unpublished	L	–	–	
Danbaekkong	PI 619083	Warrington et al. (2015)	H	+	Unknown	

[a]Study in which a significant QTL was mapped in the cqSeed protein-003 interval and the genotype was a parent. Diers et al. (in prep) refers to unpublished results in a SoyNAM population (Diers et al. 2018), Chaky refers to Chaky et al. (2003), Warrington refers to Warrington et al. (2015), Kim refers to Kim et al. (2016), and P followed by a number refers to Phansak et al. (2016) with the number corresponding to the population number in that study.

accessions, 167 had the Williams 82 allele for all five markers (haplotype 1), 53 were homozygous for the Williams 82 allele for the four flanking SNP markers but the PI 468916 allele for *Glyma.20g085100* (haplotype 2), three homozygous for the Williams 82 allele for the flanking markers and heterozygous for *Glyma.20g085100* (haplotype 3), one homozygous for the PI 468916 allele for the four flanking markers and the Williams 82 allele for *Glyma.20g085100* (haplotype 4), 13 homozygous PI 468916 for all five markers (haplotype 5), and one homozygous for PI 468916 for all but one flanking marker that was heterozygous (haplotype 6). The high number of haplotype 2 accessions show that *Glyma.20g085100* indel has unexpectedly low levels of LD with flanking markers. Additional flanking markers were used in an LD analysis, which further shows that *Glyma.20g085100* was not in strong disequilibrium with the surrounding SNPs (Figure 2a) or with the indel polymorphism in *Glyma.20g085200*, despite this gene being located only 5160 bp from *Glyma.20g085100* on the Gmax2.0 map assembly (SoyBase, https://soybase.org). Because of this lack of correlation with very closely adjacent SNPs that would normally be presumed to be strongly if not perfectly correlated to any polymorphism in the gene, the 321-bp indel is more strongly associated with the cqSeed Protein-003 trait than the other surrounding polymorphisms, explaining inconsistent association peaks found in previous association studies. The lack of correlation is explained by a lack of association between the presence of the indel and overall relatedness of the lines studied, as assessed by PCA (Figure 2b).

These results indicated that the CCT protein-encoding *Glyma.20g085100* was the most likely causal locus for cqSeed protein-003, and that the 321-bp indel was thus the likely causal polymorphism. The 321-bp indel occurs at the beginning of Exon 3 (Figure 3a), causing the annotated splice site to be altered and the size of the exon to change (Figure 3b). In total, 73 amino acids were changed in the predicted protein because of this indel, with the low-protein, Williams 82 allele having 37 additional residues as well as a number of substitutions and deletions in a conserved region of the protein (Figure 3c). By re-annotating the Williams 82 version of this gene, we were able to find another likely coding region of the genome, coding for a conserved c-terminal region of the CCT domain, after the region of the protein disrupted by the 321-bp insertion that is conserved with other CCT domain proteins but not present in the current (a2.v1) annotation. However, either version of the annotation shows that the PI 468916 version of the gene codes for a CCT domain protein highly conserved with other related proteins, but the insertion causes the Williams 82 version to code for a highly divergent protein (Figure 3c).

Interestingly, the 321 bp present in the Williams 82 allele, but not present in the PI 468916 allele, also showed strong similarity to transposable elements, including a TIR type DNA transposon (Table S5). This finding indicates that the low-protein allele found in elite *G. max* soybean may be of evolutionarily recent origin, and raises the possibility that the sequence had the potential to revert from the low-protein genotype to the high-protein *G. soja* allele by excision of the fragment. Evidence for possible reversion to the high-protein genotype is given in Figure 2b, where a number of accessions with the high-protein allele, but exhibiting an otherwise Williams 82 haplotype throughout the region (haplotype 2), were observed scattered among low-protein accessions.

Three germplasm accessions, PI 407877B, PI 423954, and PI 424148, were identified with the shorter version (PI 468916) of the *Glyma.20g085100* gene, but were previously associated with the low-protein version of the cqSeed protein-003 locus. As this was inconsistent with the interpretation that the short version of *Glyma.20g085100* was responsible for the high-protein phenotype, we evaluated this interpretation in a new population. The three

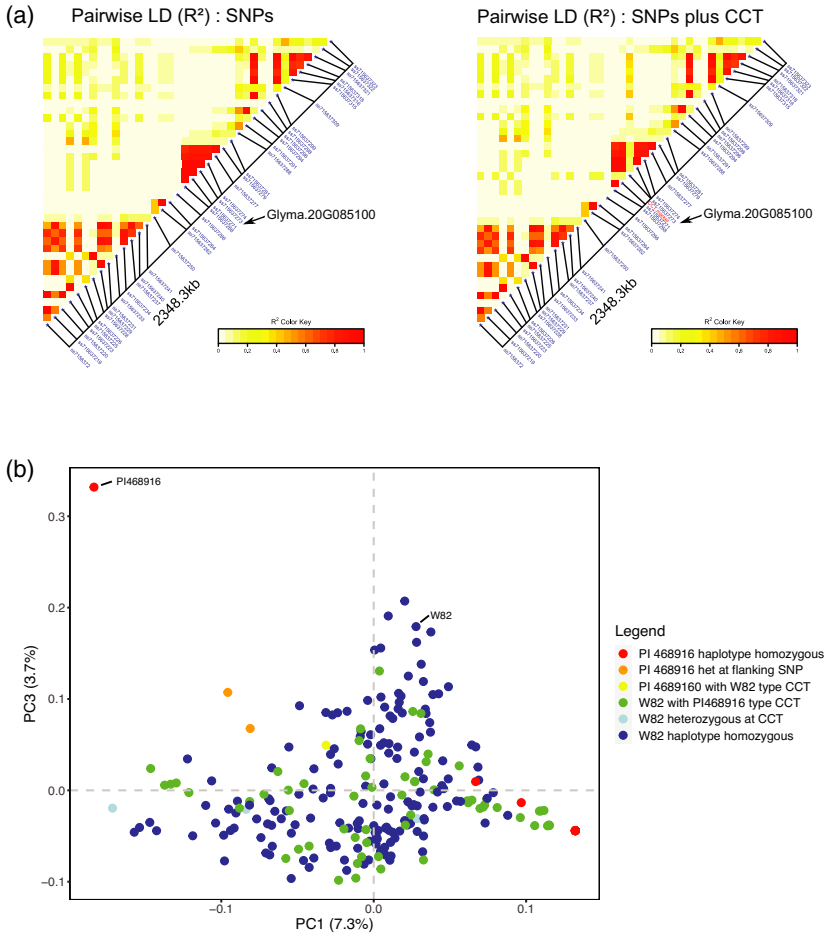

Figure 2. Population biology of the insertion/deletion (indel) polymorphism at *Glyma.20G085100*.
(a) Linkage disequilibrium (LD) around the *Glyma.20G085100* locus. Left panel: using a population of diverse accessions (Tables 3 and S4) the R^2 LD values around the locus were calculated using data from the SoySNP50k array (Song et al. 2013); note there is a region of LD surrounding the *Glyma.20G085100* gene indicated by the arrow (left panel). Right panel: indel genotype determined with a polymerase chain reaction-based marker detecting the insertional polymorphism in the CCT-domain gene (CCT marker) is also added. It is clear the indel locus is not in strong LD with surrounding markers. (b) Principal components analysis of the accessions studied here for protein content. First and third principal components were calculated from whole-genome SoySNP50k array marker information for each accession and plotted, and the points representing accessions colored by the genotype at the four single nucleotide polymorphisms (SNPs) in (a) in LD with the *Glyma.20G085100* locus plus the *Glyma.20G085100* indel marker. Accessions with the PI 468916 sequence across the locus are denoted in red and Williams 82 sequence in blue. Lines were also detected that were homozygous or heterozygous for the PI 468916 version of the indel in *Glyma.20G085100*, but carried the flanking haplotype of markers identical to Williams 82, denoted in light blue or green. Broad distribution of green points indicates likely reversion by transposon excision.

inconsistent accessions were from QTL mapping by Phansak et al. (2016), which involved relatively small populations of F_2 plants. Therefore, one explanation for this inconsistency is that the Type II error probability (failure to reject a null hypothesis of no QTL) can be quite high in these small populations. To assess this possibility, PI 423954 was crossed to the cultivar Williams 82, which is known to have the low-protein allele. An F_2 population developed from this cross was grown in the field and tested, along with the parents, for the 321-bp deletion in *Glyma.20g085100* and assayed for seed protein by near infrared reflectance (NIR) spectroscopy. A strongly significant association ($P < 0.01$) between the size polymorphism in the marker and protein content was detected, with the heterozygous individuals having an intermediate phenotype. This new evidence shows that this accession does have the high-protein allele, indicating that the failure by Phansak et al. (2016) to detect it was likely attributable to a Type II error, and suggesting that this may be occurring in the failure to detect it in the other two accessions.

Design, construction, and testing of RNAi transgenic downregulation events

To confirm that the candidate gene *Glyma.20G085100* functionally regulates seed protein content by being coincident with the causative polymorphism in cqSeed protein-003, transgenic plants were generated in an attempt to knock-down transcription of the native gene. Because of

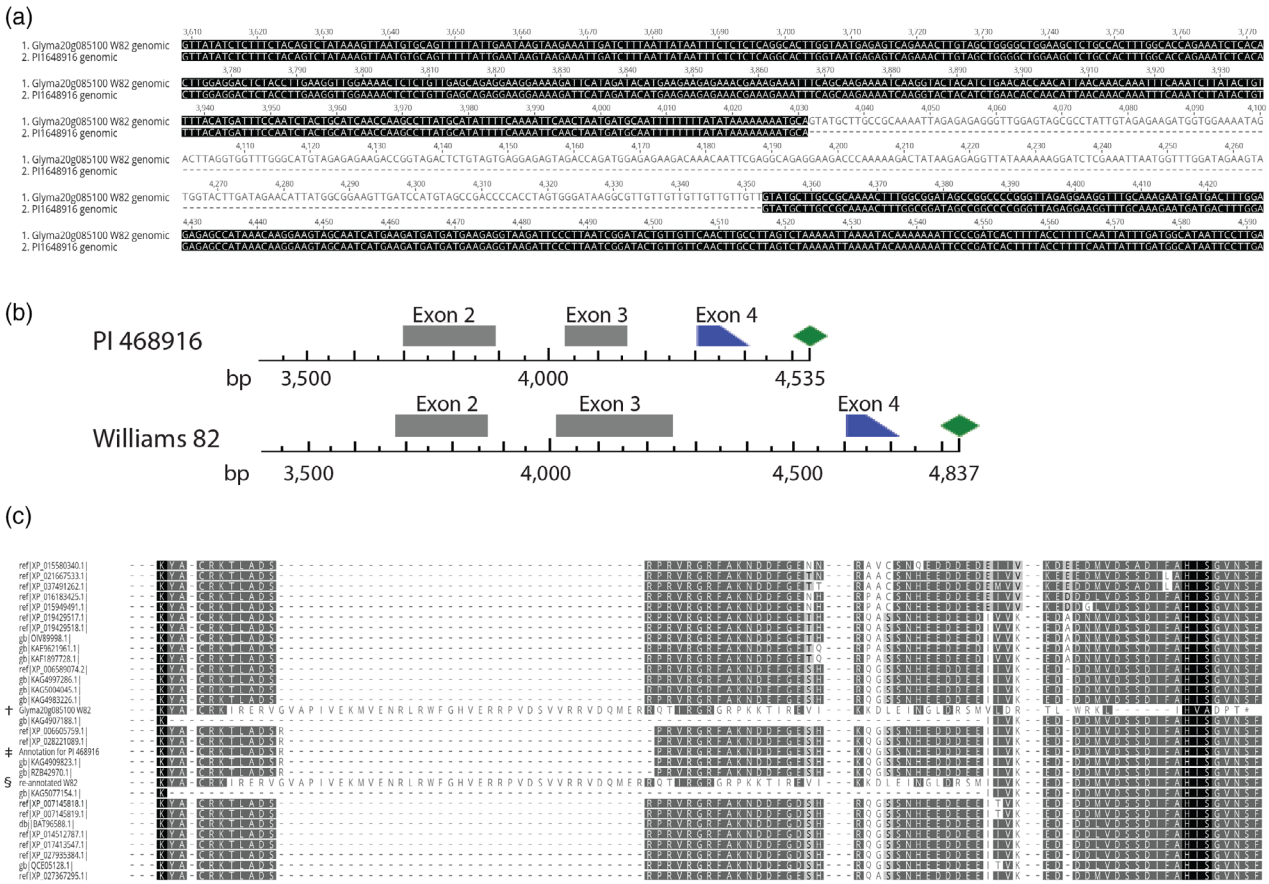

Figure 3. Sequence of the insertion/deletion polymorphism.
The 321-bp indel polymorphism in the *Glyma.20g085100* gene responsible for the cqSeed protein-003 quantitative trait loci.
(a) A 321-bp sequence with transposon similarity is present in the Williams 82 genome and not the genome of the high-protein accession PI 468916.
(b) Predicted impact of the indel shown in A on the intron–exon structure of the *Glyma.20g085100* transcript. Internal coding exons are shown as gray blocks (Exon 1 is identical and not shown), terminal exon as a blue block, and polyadenylation signal as a green diamond. Positions are in base pairs from the transcription start site, as in (A). The orange box shows the region of the insertion.
(c) Multiple alignment of related CCT domain proteins. Published protein sequence of *Glyma.20g085100* is shown (†) along with our annotation of the sequence from PI 468916 (‡) and a re-annotated version of the Williams 82 sequence using the same methods as used for the PI 468916 sequence (§) aligned to related proteins in GenBank. Note that the sequence of PI 468916 is closely conserved with related proteins.

the degree of sequence similarity within most of the protein coding region of *Glyma.20G085100* and its homeologous region on chromosome 10, and the very high similarity between the inserted region and many other parts of the genome, a 300-bp region was chosen in the 5′ UTR of *Glyma.20G085100* for the RNAi construct, which has a unique sequence with low similarity to other regions of the soybean genome. A hairpin loop construct designed to knock-down expression of the *Glyma.20G085100* gene, driven by the CaMV 35S promoter, was transformed into the soybean cultivar Thorne, which is a low-protein line with a Williams 82 type allele at *Glyma.20G085100*.

Primary transgenic events (T_0) plants were established in the greenhouse, and allowed to self-pollinate. Integration patterns across the independent events were ascertained via Southern blot analysis on selected T_1 individuals (Figure S2). Populations of T_2 plants derived from two events, 1146-5 and 1157-1, harboring one and two transgenic alleles, respectively (Figure S2), were selected for phenotypic characterization. The transgenic allele was monitored in the T_2 generation by PCR and an herbicide tolerance assay to determine the effects of the RNAi construct on the seed protein phenotypic under greenhouse conditions. Relative transcript levels of *Glyma.20G085100* in the events was monitored in leaf tissue (V5 stage of development) and in immature embryos (R5 stage of development) (Figure S3). The monitoring of T2 and T3 individuals across seven events revealed relative reductions in the *Glyma.20G085100* transcript in both V5 and R5 stages in events including 1146-5 and 1157-1 (Figure S3).

The protein content of T_3 seed harvested from the individually threshed T_2 plants was measured using NIR. Relatively few null segregants were detected, likely because the transgene was present in the population in multiple,

Table 4. Associations between RNAi hairpin transgene and seed protein percentage on a dry weight basis in g kg^{-1} for three populations of greenhouse grown T_2 plants

	Transgene present		Transgene not present		
Population	*n*[a]	Mean	*n*[a]	Mean	Pr > F*
1157-1-T1-3	18	417	5	399	0.05
1146-5-T1-4	25	444	5	415	0.04
1146-5-T1-5	20	441	5	409	0.02

[a]Number of plants tested that did or did not have the transgene present.
*Significance level of the test to detect differences between the present and not present groups.

unlinked copies. The difference in protein content between seed from plants with and without the transgenic allele was statistically significant (Table 4). For one event, the transgenic plants showed on average >3% more protein than control segregants. Thus, a knock-down construct to reduce expression of the low-protein version of the *Glyma.20G85100* gene was capable of increasing seed protein concentration. These results further support our conclusion that the 321-bp insertion/deletion variant in the CCT-protein encoding gene *Glyma.20G85100* is likely the cause of the seed protein phenotype and thus constitutes the molecular gene basis of the cqSeed protein-003.

DISCUSSION

In this study, we narrowed the location of cqSeed protein-003 to a 77.8-kb region on chromosome 20 flanked by the genetic markers BARCSOYSSR_20_0674 and BARCSOYSSR_20_0670. The physical position of the interval was between 31.74 and 31.82 Mbp based on the Gmax2.0 assembly, which is the narrowest interval that cqSeed protein-003 has been mapped to date. This narrowed region coincides with the candidate gene region identified in the most recent GWAS by Bandillo et al. (2015), where they mapped the QTL to the approximately 2.4-Mbp interval located between 30.7 and 33.1 Mbp. However, the region identified in our study does not concur with the regions assigned to this QTL in two other recent GWAS studies. Hwang et al. (2014) mapped the QTL to the 2.4-Mbp interval located at 28.7–31.1, and Vaughn et al. (2014) mapped the QTL to the approximately 1-Mbp region located at 32.1–33.1 Mbp. It is not surprising that this QTL has been difficult to map in association studies as the insertion *Glyma.20G85100* has low LD with surrounding markers (Figure 2a).

We conclude, based on both marker correlation analysis and transgenic plant studies, that the 321-bp deletion in the PI 468916 allele, relative to the Williams 82 reference allele, in *Glyma.20g085100* is the causative molecular locus for cqSeed protein-003. This gene encodes a CCT (CONSTANS, CONSTANS-like, TOC1) motif family protein, a motif that has been shown to be present in genes that control flowering in Arabidopsis (Masaki et al., 2005), in crops (Zhang et al., 2015), and it has been shown to be present in GATA-type zinc-finger proteins that act as protein recognition motifs (Liew et al., 2005). Although the Williams 82 version of the protein encodes an aberrant CCT domain, the PI 468916 version of the gene codes for a highly conserved CCT domain (Figure 3c). While strongly associated with biological timing, CCT domain proteins have also been implicated in the regulation of a number of other processes in crops, including photosynthesis, nutrient use efficiency, and stress tolerance (Liu et al., 2020). However, this gene is sufficiently divergent from the canonical sequence of the CCT motif that it was not listed in a recent review of soybean CCT domain proteins (Mengarelli and Zanor, 2021). The possible role of this gene in biological timing is of interest, in view of a possible role in the timing of grain filling as well as other processes relevant to the accumulation of protein in the soybean seed.

There is evidence that *Glyma.20g085100* impacts plant maturity in addition to protein and oil. It has been shown that lines homozygous for the allele for higher protein also matured 0–5 days earlier than lines homozygous for the alternative allele (Sebolt et al. 2000; Brzostowski et al., 2017; Prenger et al., 2019) but no literature was found that evaluated its effect on date of flowering. Effects of this gene on protein and maturity have been found in mapping populations that range from MG II, which is grown in the northern USA (Sebolt et al., 2000), to MG VII, which is grown in the southern USA (Prenger et al., 2019). For example, Brzostowski et al. (2017) evaluated two MG IV populations segregating for *Glyma.20g085100* allele and found in one population a significant maturity effect of 1 day between lines homozygous for the two different alleles and in the second population a difference of 3 days. The allelic effect on protein was almost identical between the two populations (22 and 23 g kg^{-1}). These results show that the gene impacts protein and oil across growing regions and genetic backgrounds and that these effects are not necessarily associated with large differences in maturity. Any effect that was largely mediated by maturity would be expected to show large genotype × environment effects across different growing regions, leading us to speculate that the effect on seed protein level of the CCT protein encoded at *Glyma.20g085100* could be related to another regulatory system such as the regulation and sensing of sugar concentration (Masaki et al., 2005). Further research is needed to improve our understanding of how *Glyma.20g085100* affects plant maturity and how this is related to seed composition.

The PI 468916 sequence of *Glyma.20g085100* is shorter than the sequence from Williams 82, which has the lower protein allele for the QTL (Kim et al., 2016). Interestingly,

the deletion/insertion has low LD with other SNPs and markers in the region (Figure 2), and the 321-bp indel shows similarity with TIR-type DNA transposons (Table S5), indicating that the most likely explanation is that the mutation is actually a transposon insertion in the low-protein lineage. In support of this hypothesis, sequenced wild (i.e., *G. soja*) soybeans and their relatives with available sequences at the time of writing, were found to have the high-protein, PI 468916 version of this gene, indicating that the low-protein allele is a relatively recent mutation that may have been selected within elite (*G. max*) germplasm in the past. In addition, the low-protein version appears to revert to the high-protein genotype, as evidenced by the presence of many high-protein genotypes with completely conserved low-protein flanking haplotypes (haplotype 2) among genetically similar low-protein accessions (Figure 2b). These revertants are likely to have been generated by excision events. Finally, the high-protein version of the gene contains a well-conserved CCT domain with many other related genes (Figure 3c), giving further evidence that it is the older version of the sequence.

We found that by reducing expression of the *Glyma.20g085100* gene using RNAi, we were able to increase protein levels in the Thorne (Williams 82 type, low protein) genetic background. This result implies that this version of the protein encoded by the allele with the 321 bp insertion is functional, and reducing its expression using RNAi, and thus presumably the amount and activity of its protein product, increases protein levels. Unusually for a gain-of-function mutation, this allele affects the protein-coding region rather than the regulatory region of the gene. Transposons are known to cause exon shuffling (Quesneville, 2020), which may have contributed to a gain of function in the protein encoded at this locus, through disruption of a major part of the CCT domain (Figure 3c). We conclude that the insertion may result in a novel function that causes the reallocation of resources away from seed protein and into carbohydrates and lipids analogous to the other major protein QTL in soybean on chromosome 15 that was recently cloned (Wang et al., 2020). A gain of molecular function is consistent with the significant alteration and extension of the protein sequence encoded by the gene caused by the insertion (Figure 3).

In conclusion, we fine mapped cqSeed protein-003 to a 77.8-kb region on chromosome 20 between the SSR markers BARCSOYSSR_20_0674 and BARCSOYSSR_20_0670, with each corresponding to the physical positions of 31 821 947 bp to 31 744 150 bp, respectively, based on the Gmax2.0 map assembly. This narrowed candidate gene region allowed us to identify the causal allele for the most widely studied, large effect protein QTL in soybean. The allele itself appears to be the result of a transposon insertion in the *Glyma.20g085100* gene, leading to a low-protein allele at this locus, which seems to have reverted to the high-protein allele in some accessions. The gene encodes a CCT domain protein, encoding an aberrant CCT domain in the Williams 82 allele, possibly involved in the biological timing of seed biogenesis and development.

Soybean breeders have struggled with balancing breeding for increased seed protein content with its negative correlations with yield and oil. Breeders largely focus on increasing yield in their breeding programs and this has resulted in a commensurate reduction in protein content. Rincker et al. (2014) showed that over 80 years of breeding mostly focused on increasing yield that protein content decreased by about 2% or 20 g kg^{-1} seed. Zhang et al. (2020) recently reported the identification of a polymorphism in the GmSWEET39 gene, which was strongly associated with protein and oil content in soybean seed and is potentially the basis of the important chromosome 15 protein QTL. The identification of these genes that control seed protein and oil content may open new approaches for modifying these genes in ways that can result in increases in protein without a significant reduction in yield and oil.

EXPERIMENTAL PROCEDURES

Fine mapping

The cqSeed protein-003 locus was fine mapped through an iterative process of developing and testing backcross populations segregating for different sections of the region it maps. PI 468916 is the source of the high-protein allele and this allele was backcrossed into the background of A81-356022, an Iowa State University experimental line. Populations used in the current study were developed from backcross lines described by Nichols et al. (2006).

Below is a brief explanation of the development and testing of germplasm used in the fine mapping. See Methods S1 for a detailed explanation of the process and Figure S1 for an annualized outline of the steps taken in the mapping. The fine mapping was done in two rounds with the second round efforts focused on the QTL interval identified in round 1. Across the two rounds, 7603 plants segregating for the QTL in backcross populations were screened with markers that flanked the interval the QTL maps to identify plants with crossovers between the markers. The selected crossover plants were then tested with all available markers (Bolon et al., 2010; Song et al., 2010) in the QTL interval to map the location of the crossovers. Plants with crossovers that mapped to unique locations were selected and a population of plants or lines were then developed from each selected plant and grown in the field and tested with a segregating marker using methods from Cregan and Quigley (1997), Keim and Shoemaker (1988), and Wang et al. (2003). Seed harvested from the plants and lines were tested for protein and oil content using near infrared transmittance. The marker and composition results from each population were then analyzed using the PROC GLM function of SAS (SAS, 2016). This analysis of each population resulted in a narrowing of the QTL interval as a significant association indicated that the QTL was within the interval segregating in the population with no significance indicated that it is not within the segregating interval.

Candidate gene evaluation

To test candidate genes in the fine-mapped cqSeed protein-003 interval, a panel of 53 genotypes was assembled. These genotypes had served as high- or low-protein parents of multiple biparental populations that had been used in seed protein and oil QTL mapping studies in which the allelic segregation of a cqSeed protein-003 QTL was inferred (Table 3). Of the 53 mapping population parents, 32 were documented as homozygous for a high-protein allele in the cqSeed protein-003 interval, whereas there were 21 homozygous for the low-protein allele. DNA was extracted from these parents and tested with markers developed from polymorphisms within *Glyma.20G85100* and *Glyma.20G85200*, two candidate genes in the interval that cqSeed protein-003 was mapped. For *Glyma.20G85100*, a codominant marker was designed giving bands with a 321-bp size differential between the two alleles using the primer sequences ACTGCATCAACCAAGCCTTATGC and TGTACGTTTCTAACTCACTTAACTTATTGG. For the much larger size difference in *Glyma.20G85200*, it was necessary to use two individual dominant markers. The primer sequences used for the longer allelic variant were CATGGGTAGTTTCTGAAAGCA and CGAGTCTTTCAAAGCATACCA, and for the shorter variant, the primer sequences were TAGTGTCTACTGTACGTAAGTT and CGATATCCAAGTGAACGC. Together, the data collected from the longer and shorter variant gave codominant marker results.

The candidate gene marker analysis was conducted using a 20-μl PCR protocol containing 1 μl of 5 μM forward and reverse primers, along with the components from the TaKaRa DNA polymerase kit (TBSUSA, Mountain View, CA, USA): 0.1 μl TaKaRa ExTaq DNA polymerase, 1.6 μl dNTP, and 2 μl 10× buffer with 1 μl of a five-fold diluted template DNA. PCR amplicons were visualized on a 1% agarose electrophoresis gel run at 100 V for 30 min (Bio-Rad Hercules, CA, USA).

De novo sequence assembly and analysis

Whole genome sequencing of the high-protein accession PI 468916 was performed using the HiSeq v.1 DNA sequencer (Illumina, San Diego, CA, USA) at the Carver Biotechnology Center at the University of Illinois. DNA was extracted using the method described above. The raw reads were assembled using ABySS (Simpson et al., 2009) to generate scaffolds. In total, 182 255 190 paired end reads were assembled, each 150 nucleotides in length, with an average gDNA fragment size of 580 nucleotides. A *k-mer* of 64 was chosen as the best performing (ABySS was compiled for a maximum *k-mer* size of 64). The Williams 82 reference sequence (a2.v1) corresponding to the region between the flanking markers, was compared with the scaffold output from the assembly of PI 468916 sequences using the BLAST program to locate potential scaffolds corresponding to the targeted QTL interval. These scaffolds were then loaded into the DNA analysis software GENEIOUS 11.1.5 (Geneious, Auckland, New Zealand). The sequences from PI 468916 were mapped to the reference Williams 82 and examined for single nucleotide variation and structural changes in genic regions and aligned against the Williams 82 reference. All discrepancies were noted and investigated for potential function, although only three polymorphisms were found that affected protein-coding genes within the final marker-flanked QTL interval.

LD analysis

LD was calculated as a pairwise R^2 value from SNP markers on either side of the *Glyma.20G85100* marker for the panel of parents described above using the R package genetics (Warnes, 2003). The SNP data for the panel were generated using the SoySNP50k array to genotype the soybean germplasm collection (Song et al., 2013) (data available at https://soybase.org/snps/). These R^2 values were visualized as a heat map and plotted to show the relationship of each SNP to each other SNP using the R package LDheatmap.

PCA and haplotype analysis

PCA analysis was performed and visualized using SNPrelate package in Bioconductor for R (Zheng et al., 2012). SNP data for the population of accessions was generated from the SoySNP50k SNP data (Song et al., 2013) (data available at https://soybase.org/snps/). Details of accessions including genotypes of the extracted SNPs and indel marker are given in Tables 2 and S4.

Haplotypes were determined for the *Glyma.20G85100* interval using the codominant marker for this gene and the two closest markers (ss715637268 and ss715637271) flanking one side of the indel in the gene and two on the other side (ss715637273 and ss715637274). The haplotypes were determined for these five markers for the soybean accessions listed in Tables 3 and S4 and were the same as those in Figure 2b, coded as follows for the tables: 1 = Williams 82-like across all five markers; 2 = Williams 82-like for the four SNP markers and PI 468916-like for the *Glyma.20G085100* marker; 3 = Williams 82-like for the four SNP markers and heterozygous for the *Glyma.20G085100* marker; 4 = PI 468916-like for the other four SNP markers Williams-like for the *Glyma.20G085100* marker; 5 = PI 468916-like for all five markers; and 6 = PI 468916-like for all markers except heterozygous for ss715637268.

RNAi vector construction and soybean transformation

A hairpin element was assembled using a 299-bp sequence from the Williams 82 genome version a2.v1 beginning at position 31 774 770 and ending at 31775069 on Chr 20, and representing most of the 5′ UTR of *Glyma.20G85100*. The element was synthesized (GenScript USA, Piscataway, NJ, USA), with the incorporation of *Spe*I and *Bgl*II sites on the ends, to allow inverted elements to be subcloned into the vector pUC58-RNAi (gift from H. Cerutti University of Nebraska) (Figure S4). The resultant inverted elements were separated by the second intron of a small nucleolar protein from Arabidopsis (At 004G02840) to facilitate hairpin folding (Wesley et al., 2001). The resultant element was subcloned between the 35S CaMV promoter and T35S terminator from cauliflower mosaic virus. The subsequent RNAi expression cassette was then cloned into a binary vector that harbored a bar gene plant selection cassette and the final binary vector for soybean transformation designated pPTN1379 (Thompson et al., 1987) (Figure S4). The binary vector pPTN1379 was mobilized into *Agrobacterium tumefaciens* strain EHA101 (Hood et al., 1986) by triparental mating and the derived transconjugant was used to transform the soybean cultivar Thorne (McBlain et al., 1993) following the protocol previously communicated (Zhang et al., 1999).

Southern blot analysis

Southern blot analysis was conducted as previously described (Eckert et al., 2006). Ten micrograms of total genomic DNA was restriction digested with *Eco*RI and separated on a 0.8% agarose gel. The separated DNAs were transferred to a nylon membrane and hybridized with dCT ^{32}P-labeled 5′ UTR region used in the hairpin design.

Transgenic plant materials and DNA extraction for PCR genotyping

Transgenic seeds were surface sterilized in 10% bleach and sown in germination paper. Germinated disease-free seedlings were transplanted into soil-filled 30 cm diameter pots in a greenhouse. Plants were grown under a 14.75-h day length with temperatures during the day that ranged from 28 to 30 C and night from 23 to 26 C. DNA extraction was performed on young trifoliate leaf tissue using a modified protocol from Keim and Shoemaker (1988).

Transgenic plant growth, genotyping, and protein analysis

Three populations of T_2 seed from two T_0 transformation events were developed and grown in a greenhouse. Two T_2 populations (T_1-4 and T_1-5) were developed from the T_1 event 1146–5 and the other population (T_1-3) was from event 1157–1. These T_2 lineages were grown in the greenhouse under the conditions described above and DNA was extracted according to Keim and Shoemaker (1988). These plants were genotyped to determine the presence of the transgenic allele in each plant. Presence of the transgenic allele was tested using a PCR marker designed with primers inside the transgene construct. The primers had the sequences 5′-CGAGGAGGTTTCCGGATATTAC-3′, and 5′-GCAC GACACACTTGTCTACT-3′; PCR conditions were 1 min initial incubation at 95°C, followed by 30 cycles of 30 sec at 95°C, 30 sec at 52°C, 1 min at 72°C, followed by 65 min at 72°C final extension. The marker detected the presence of the transgene and gave a dominant phenotype and therefore was unable to differentiate whether the gene was present in the homozygous or heterozygous state. Seed from each plant was threshed individually, and protein and oil content was measured with an NIR (Perten DA 7250, Stockholm, Sweden). Seeds that had visible mold, which had begun to germinate or were damaged, were excluded from the measurement. The seeds were repacked and protein content was measured a second time and averaged.

Selectable marker analysis of transgenic plants

The segregation of the transgenic allele in progeny of the T_2 lineages was also monitored using herbicide tolerance assay (Zhang et al., 1999). For herbicide resistance selection, a 100× dilution of Finale, a commercial formulation of glufosinate was used (BASF, Ludwigshafen, Germany). A 0.25% (v/v) of a non-ionic surfactant (Activator 90; Loveland Products Inc., Loveland, CO, USA) was added to enhance leaf wetting. First trifoliate leaves were painted with the herbicide across the leaves. Bands of leaf chlorosis and necrosis were scored after 5 days.

Reverse transcription-PCR analysis of RNAi events to monitor expression changes in *Glyma.20G85100*

As a means to assess relative transcript changes between the RNAi events and wild-type Thorne plants, total RNA was isolated from vegetative tissue at the V5 stage of development and immature embryos at the R5 stage of development, under field conditions. Total RNA was extracted with the RNeasy Plant Mini Kit (QIAGEN Corp. cat. no./ID: 74904, Hilden, Germany) according to the manufacturer's instructions. Two micrograms of total RNA were used for cDNA synthesis with the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific, Waltham, MA, USA). PCR was conducted with a thermo cycle 94°C/1 min and 30 cycles of 94°C/30 sec, 55°C/30 sec, and 72°C/30 sec followed by an additional 72°C/5 min for extension. Primers Gm20P_forward (5′-TCCACCACTTCTCCCAATCTCAAC-3′) and Gm20P_reverse (5′-CCCGTCAAAATTGAACCTGCTG-3′) to amplify Gm20P specific fragments and Gm-Actin6_L2 (5′-TGGTGTGATGGTTGGCATGG-3′) and Gm-Actin6_R2 (5′-GGGTTAAGAGGGGCCTCAGT-3′) to amplify *GmActin6* fragments as controls using GoTaq® Master Mixes (Promega Corporation, Madison, WI, USA).

ACKNOWLEDGEMENTS

This work was partially supported by soybean check-off funding from the United Soybean Board and the North Central Soybean Research Program and by the USDA National Institute of Food and Agriculture, Hatch, project 1015014.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

The whole-genome DNA sequence scaffolds for PI 468916 are in the National Center for Biotechnology Information GenBank database under accession numbers MZ956137 through MZ956151.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Outline of the activities conducted on an annual basis for the fine mapping the high-protein QTL cqSeed protein-003.

Figure S2. Southern blot analysis on T_1 individuals derived from four transgenic pPTN1379 events. Ten micrograms of total genomic DNA isolated from progeny of pPTN1379 transgenic soybean events was digested with the restriction enzyme *Eco*R1. There is one *Eco*R1 site within the T-DNA element of pPTN1379. The processed membrane was hybridized with 32P labeled 299 bp element used in the RNAi hairpin. Lane 1, Thorne, refers to wild-type control. Lanes 2–8 are T_1 derived from events 1146-5, 1146-6, 1157-1, and 1158-1. Lanes with * indicate lineages used for phenotypic analyses. + control lane refers to 20 pg of free pPTN1379 plasmid cut with *EcoR*1 (linearized). Approximate 1.9-kb hybridization signal reflects the endogenous locus.

Figure S3. Relative *Glyma.20G85100* transcript changes in RNAi events observed in leaf and immature embryos S3A: diagrammatic representation of the *Glyma.20G85100* locus, highlighting location of primer annealing used in RT-PCR reaction. S3B: RT-PCR results across seven independent events carrying the RNAi element targeting 5′ UTR region of *Glyma.20G85100* gene call. V5 results from leaf tissue collected at V5 stage of development. R5 results from immature embryo tissues collected at R5 stage of development.

Figure S4. Vector pPTN1379 is a derivative of pPZP200. Expression of the RNAi element is regulated by the 35S CaMV promoter (35S promoter) and terminated by the CaMV termination sequence (T35s polyAAA). Expression of the selectable marker (bar) is controlled by the *Agrobacterium tumefaciens* nopaline synthase promoter (Pnos) and terminated by its polyadenylation element (Tnos). RB and LB, refer to the right and left border element of the T-DNA respectively. aadA refers to the bacterial selection marker for spectinomycin resistance. The ori and bom, indicate location of the origin of replications and basis of mobility, respectively, while the sta and rep regions indicate the position of the replication and stability regions of the plasmid.

Appendix S1. Supplemental description of the fine mapping methods.

Table S1. Mean protein concentration and marker-protein association (MPA) values for 13 populations of BC_5F_7 plants and their descendant $BC_5F_{7:8}$ lines tested in 2008 and 2009 during the first round of fine mapping that targeted the cqSeed protein-003 QTL on soybean chromosome 20. Marker associations based the marker listed in the marker name column.

Table S2. Mean protein concentration and marker-protein association (MPA) values for 11 populations tested during the second round of fine mapping that targeted the cqSeed protein-003 QTL on soybean chromosome 20. Populations of BC_5F_9 plants were tested in 2011, and their descendant $BC_5F_{9:10}$ lines were tested in 2012, and $BC_5F_{9:11}$ lines 2013. Mean protein content (g kg^{-1} on a 13% moisture basis) of plants and lines that were homozygous for the donor parent (*Glycine soja*) high-protein allele (B), homozygous for the recurrent parent (*Glycine max*) low-protein allele (A), or heterozygous/heterogeneous (H) based on the marker listed in the marker name column.

Table S3. Marker associations of BARCSOYSSR_20_0655 in the ProI-7 subpopulations and their mean protein content

Table S4. *Glycine max* accessions screened with *Glyma.20G085100* marker (CCT). *Glycine max accessions* were screened with the *Glyma.20G085100* (CCT) marker to determine if they had the PI 468916 genotype. *Glycine max* accessions with the PI 468916-like allele were coded as a 0. Accessions that have the Williams 82-like allele were coded with a 2. Heterozygous samples were coded as a 1. Samples that either failed in the PCR reaction or did not have DNA due to extraction or germination issues are coded as a '-'. Protein and oil content information is from the GRIN-Global website (https://npgsweb.3ars-grin.gov/gringlobal/search). Haplotypes were determined with the *Glyma.20G085100* marker and two markers (ss715637268 and ss715637271) on one side and two markers on the other side (ss715637273 and ss715637274) of the gene. Haplotypes were determined using the marker from *Glyma.20G085100* and the two closest markers on each side of the gene. The haplotype designations are defined in the Experimental Procedures section.

Table S5. Soybean transposon sequences that contain a 284 bp block of 99–100% nucleotide identity to the inserted sequence in the low-protein allele of *Glyma.20G85100*.

REFERENCES

American Soybean Association (2019) 2019 SoyStats Available at https://soygrowers.com/wp-content/uploads/2019/10/Soy-Stats-2019_FNL-Web.pdf (Verified 1 June 2021).

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J. et al. (2015) A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*, **8**, 1–13.

Bolon, Y.T., Joseph, B., Cannon, S.B., Graham, M.A., Diers, B.W., Farmer, A.D. et al. (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL. *BMC Plant Biology*, **10**, 41.

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

Brzostowski, L.F., Pruski, T.I., Specht, J.E. & Diers, B.W. (2017) Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. *Theoretical and Applied Genetics*, **130**, 2315–2326.

Burton, J.W. (1985) Breeding soybeans for improved oil quantity and quality. In: Shibles, R. (Ed.) *World soybean research conference III: proceedings*. Boulder, CO: Westview Press, pp. 361–367.

Butler, K.J., Fliege, C., Zapotocny, R., Diers, B., Hudson, M., and Bent, A.F. (2021) Soybean cyst nematode resistance quantitative trait locus cqSCN-006 alters the expression of a γ-SNAP protein. *Molecular Plant-Microbe Interactions*, **34**, 1433–1445.

Chaky, J.M., Specht, J.E. & Cregan, P.B. (2003) Advanced backcross QTL analysis in a mating between *Glycine max* and *Glycine soja* [abstract]. *Plant and Animal Genome Abstracts*, **P545**.

Chung, J., Babka, H.L., Graef, G.L., Staswick, P.E., Lee, D.J., Cregan, P.B. et al. (2003) The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Science*, **43**, 1053–1067.

Cregan, P.B. & Quigley, C.V. (1997) Simple sequence repeat DNA marker analysis. In: Caetano-Anolles, G. & Gresshoff, P.M. (Eds.) *DNA markers: Protocols, applications and overviews*. New York: John Wiley & Sons, pp. 173–185.

Cromwell, G.L. (2012) Soybean meal- an exceptional protein source. Available at http://www.soymeal.org/ReviewPapers/SBMExceptionalProteinSource.pdf.

Csanádi, G., Vollmann, J., Stift, G. & Lelley, T. (2001) Seed quality QTLs identified in a molecular map of early maturing soybean. *Theoretical and Applied Genetics*, **103**, 912–919.

Diers, B.W., Keim, P., Fehr, W.R. & Shoemaker, R.C. (1992) RFLP analysis of soybean seed protein and oil content. *Theoretical and Applied Genetics*, **83**, 608–612.

Diers, B.W., Specht, J., Rainey, K.M., Cregan, P., Song, Q., Ramasubramanian, V. et al. (2018) Genetic architecture of soybean yield and agronomic traits. *G3: Genes, Genomes, Genetics*, **8**, 3367–3375.

Eckert, H., LaVallee, B., Schweiger, B.J., Kinney, A.J., Cahoon, E.B. & Clemente, T. (2006) Co-expression of the borage Delta 6 desaturase and the Arabidopsis Delta 15 desaturase results in high accumulation of stearidonic acid in the seeds of transgenic soybean. *Planta*, **224**, 1050–1057.

Grant, D., Nelson, R.T., Cannon, S.B., and Shoemaker, R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research*, **38** (Database issue), D843–D846.

Hood, E.E., Helmer, G.L., Fraley, R.T. & Chilton, M.-D. (1986) The hypervirulence of Agrobacterium tumefaciens A281 is encoded in a region of pTiBo542 outside of T-DNA. *Journal of Bacteriology*, **168**, 1291–1301.

Hwang, E.Y., Song, Q., Jia, G., Specht, J.E., Hyten, D.L., Costa, J. et al. (2014) A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics*, **15**, 1. https://doi.org/10.1186/1471-2164-15-1.

Keim, P. & Shoemaker, R.C. (1988) A rapid protocol for isolating soybean DNA. *Soybean Genetics Newsletter*, **15**, 150–152.

Kim, M., Schultz, S., Nelson, R.L. & Diers, B.W. (2016) Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. *Crop Science*, **56**, 219–225.

Liew, C.K., Simpson, R.J.Y., Kwan, A.H.Y., Crofts, L.A., Loughlin, F.E., Matthews, J.M. et al. (2005) Zinc fingers as protein recognition motifs: structural basis for the GATA-1/Friend of GATA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 583–588.

Liu, K. (1997) *Soybeans: chemistry, technology, and utilization*. Noew York: Chapman & Hall.

Liu, H., Zhou, X., Li, Q., Wang, L. & Xing, Y. (2020) CCT domain-containing genes in cereal crops: flowering time and beyond. *Theoretical and Applied Genetics*, **133**, 1385–1396. https://doi.org/10.1007/s00122-020-03554-8.

Lu, W., Wen, Z., Li, H., Yuan, D., Li, J., Zhang, H. et al. (2012) Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. *Theoretical and Applied Genetics*, **128**, 425–433.

Masaki, T., Tsukagoshi, H., Mitsui, N., Nishii, T., Hattori, T., Morikami, A. et al. (2005) Activation tagging of a gene for a protein with novel class of CCT-domain activates expression of a subset of sugar-inducible genes in Arabidopsis thaliana. *The Plant Journal*, **43**, 142–152.

McBlain, B.A., Fioritto, R.J., St. Martin, S.K., Calip-Dubois, A.J., Schmitthenner, A.F., Cooper, R.L. et al. (1993) Registration of 'Thorne' soybean. *Crop Science*, **33**, 1406.

Mengarelli, D.A. & Zanor, M.I. (2021) Genome-wide characterization and analysis of the CCT motif family genes in soybean (*Glycine max*). *Planta*, **253**, 15.

Nichols, D.M., Glover, K.D., Carlson, S.R., Specht, J.E. & Diers, B.W. (2006) Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. *Crop Science*, **46**, 834–839.

Phansak, P., Soonsuwon, W., Hyten, D.L., Song, Q., Cregan, P.B., Graef, G.L. et al. (2016) Multi-population selective genotyping to identify soybean (*Glycine max* (L.) Merr.) seed protein and oil QTLs. *G3-Genes Genom. Genetics*, **6**, 1635–1648.

Prenger, E.M., Yates, J., Rouf Mian, M.A., Buckley, B., Boerma, H.R. & Li, Z. (2019) Introgression of a high protein allele into an elite soybean cultivar

results in a high-protein near-isogenic line with yield parity. *Crop Science*, **59**, 2498–2508.

Quesneville, H. (2020) Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mobile DNA*, **11**, 28.

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

Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S.R. et al. (2014) Genetic improvement of soybean in maturity groups II, III, and IV. *Crop Science*, **54**, 1–14.

Salvi, S. & Tuberosa, R. (2007) Cloning QTLs in plants. In: Varshney, R. & Tuberosa, R. (Eds.) *Genomics-Assisted Crop Improvement*. Netherlands: Springer, pp. 207–225.

SAS Institute. (2016) The SAS system for Microsoft Windows. Release 9.4, Cary, SAS Inst.

Sebolt, A.M., Shoemaker, R.C. & Diers, B.W. (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Science*, **40**, 1438–1444.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. & Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123. https://doi.org/10.1101/gr.089532.108.

Song, Q., Jia, G., Zhu, Y., Grant, D., Nelson, R.T., Hwang, E.-Y. et al. (2010) Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1.0) in soybean. *Crop Science*, **50**, 1950–1960.

Song, Q., Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L. et al. (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One*, **8**, E54985. https://doi.org/10.1371/journal.pone.0054985.

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. *Breeding Science*, **53**, 133–140.e.

Thompson, C.J., Movva, N.R., Tizard, R., Crameri, R., Davies, J.E., Lauwereys, M. et al. (1987) Characterization of the herbicide-resistance gene *bar* from *Streptomyces hygroscopicus*. *EMBO*, **6**, 2519–2523.

Vaughn, J.N., Nelson, R.L., Song, Q., Cregan, P.B. & Li, Z. (2014) The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3-Genes Genomes Genetics*, **4**, 2283–2294.

Wang, D., Shi, J., Carlson, S.R., Cregan, P.B., Ward, R.W. & Diers, B.W. (2003) A low-cost, high-throughput polyacrylamide gel electrophoresis system for genotyping with microsatellite DNA markers. *Crop Science*, **43**, 1828–1832.

Wang, X., Jiang, G., Green, M., Scott, R., Song, Q., Hyten, D., Cregan, P. (2014) Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. *Molecular Genetics and Genomics*, 2014, **289**, 935–949

Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Liu, Z. et al. (2020) Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. *National Science Review*, **7**, 1776–1786.

Warnes, G.R. (2003). The genetics package. *R News*, **3**, 9–13. Retrieved from https://ci.nii.ac.jp/naid/10030730040/#cit

Warrington, C., Abdel-Haleem, H., Hyten, D., Cregan, P., Orf, J., Killam, A. et al. (2015) QTL for seed protein and amino acids in the Benning Danbaekkong soybean population. *Theoretical and Applied Genetics*, **128**, 839–850.

Wesley, S.V., Helliwell, C.A., Smith, N.A., Wang, M., Rouse, D.T., Liu, Q. et al. (2001) Construct design for efficient, effective and high-throughput gene silencing in plants. *The Plant Journal*, **27**, 581–590.

Wilcox, J.R. (1985) Breeding soybeans for improved oil quantity and quality. In: Shibles, R. (Ed.) *World soybean research conference III: proceedings*. Boulder: Westview Press, pp. 380–386.

Zhang, Z.Y., Xing, A.Q., Staswick, P. & Clemente, T.E. (1999) The use of glufosinate as a selective agent in Agrobacterium-mediated transformation of soybean. *Plant Cell Tiss. Org.*, **56**, 37–46.

Zhang, L., Li, Q., Dong, H., He, Q., Liang, L., Tan, C. et al. (2015) Three CCT domain-containing genes were identified to regulate heading date by candidate gene-based association mapping and transformation in rice. *Scientific Reports*, **5**, 7663.

Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M.L. et al. (2020) Selection of GmSWEET39 for oil and protein improvement in soybean. *PLoS Genetics*, **16**, e1009114. https://doi.org/10.1371/journal.pgen.1009114.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C. & Weir, B.S. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328. https://doi.org/10.1093/bioinformatics/bts606.