

Article

Identification of Wild Segments Related to High Seed Protein Content Under Multiple Environments and Analysis of Its Candidate Genes in Soybean

Ning Li, Mengdan Cai , Wei Luo, Wei Han, Cheng Liu, Jianbo He , Fangdong Liu, Lei Sun, Guangnan Xing , Junyi Gai * and Wubin Wang *

Sanya Institute of Nanjing Agricultural University, Jiangsu Key Laboratory of Soybean Biotech-Nology and Intelligent Breeding, Soybean Research Institute, National Innovation Platform for Soybean Breeding and Industry-Education Integration, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing 210095, China

* Correspondence: sri@njau.edu.cn (J.G.); wangbean@njau.edu.cn (W.W.)

Abstract

Annual wild soybean is characterized by a high protein content. To elucidate the genetic basis, this study utilized a chromosome segment substitution line population (177 lines) constructed with cultivated soybean *NN1138-2* as the recipient and wild soybean *N24852* as the donor. Phenotypic analyses across three environments revealed significant variation in protein content ranging from 42.86% to 49.08%, with a high heritability of 0.70, indicating strong genetic control. Through high-throughput sequencing, six wild segments associated with high protein content were detected on chromosomes 3, 6, 9, 15, and 20, with phenotypic variation explained (PVE) by individual segments ranged from 3.58% to 22.46%, with segments on chromosomes 9, 15, and 20 as large-effect segments with PVE > 10%. All wild segments exhibited positive additive effects (0.42–1.09%), consistent with the characteristic of a high protein content in wild soybean. Compared with previous studies, five segments overlapped with reported loci, while *qPro6.1* on chromosome 6 was a novel discovery. Integration of genomic and transcriptomic data identified 10 genes involved in nucleic acid binding, transmembrane protein transport, and amino acid synthesis pathway, with homologs validated in soybean, rice, and rapeseed. This research deepens the understanding of wild soybean's high protein and offers new gene resources for breeding high-protein cultivated soybean.

Keywords: wild soybean; chromosome segment substitution line; protein content; QTL; candidate gene



Academic Editor: Sang-Won Lee

Received: 24 November 2025

Revised: 14 December 2025

Accepted: 15 December 2025

Published: 17 December 2025

Citation: Li, N.; Cai, M.; Luo, W.; Han, W.; Liu, C.; He, J.; Liu, F.; Sun, L.; Xing, G.; Gai, J.; et al. Identification of Wild Segments Related to High Seed Protein Content Under Multiple Environments and Analysis of Its Candidate Genes in Soybean.

Agronomy **2025**, *15*, 2902. <https://doi.org/10.3390/agronomy15122902>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Soybean (*Glycine max*) is an important food and economic crop. Its seed protein content is approximately 40%, about 4–5 times that of other cereals such as wheat and corn, accounting for over 50% of the global plant protein supply [1]. Annual wild soybean (*Glycine soja*) is the wild ancestor of cultivated soybean and possesses the typical characteristic of a high protein content. Therefore, clarifying the genetic basis of the high protein content of *Glycine soja* is of great significance for improving cultivated soybeans [2]. However, when utilizing *Glycine soja*, the hybrid offspring often exhibit “linkage drag”, where the introduction of high-protein genes is accompanied by undesirable traits such as vining growth and small seeds, leading to low breeding utilization efficiency of superior wild

genes [3]. Therefore, deciphering the genetic basis of the high protein content in *Glycine soja* and developing molecular markers tightly linked to high protein content are the core prerequisites for breaking the “favorable gene–unfavorable trait” linkage and achieving efficient utilization of wild resources.

Soybean protein content is a complex quantitative trait controlled by multiple genes/QTL (quantitative trait loci) and is easily influenced by the environment [4]. The methods of evaluating soybean protein content mainly include the Kjeldahl nitrogen determination method and near-infrared spectroscopy [5]. The former has issues such as complex operation and long detection time, while the latter, although convenient and fast, has problems like high model requirements and low detection sensitivity. In recent years, with the rapid development of DNA molecular marker technology, many QTL for soybean protein content have been identified, enabling the possibility of marker-assisted selection. Up to 2025, the database of Soybase (<http://www.soybase.org>) has recorded 255 QTL related to soybean seed protein content, covering all 20 soybean chromosomes, but only 30 of these QTL were derived in *Glycine soja* [6–10], accounting for less than 12% of the total. Furthermore, these QTL were mostly mapped based on segregating populations like recombinant inbred lines and F₂, which suffer from genetic complexity and high false positives, resulting in mapping data which are difficult to directly utilize in molecular breeding.

Chromosome segment substitution lines (CSSLs) are ideal materials for QTL identification, fine mapping, map-based cloning, and QTL interaction studies, and have been successfully applied in plants such as rice and tomato [11–15]. To explore the favorable alleles deposited in the wild soybean, our research group, using the elite variety *Nannong 1138-2* (NN1138-2) as the recipient and wild soybean N24852 as the donor, firstly constructed a chromosome segment substitution line (CSSL) population *SojaCSSLP1* covering the entire genome of wild soybean through two to five backcrosses, more than two selfings, and marker-assisted selection over two generations. We elucidated the genetic basis of flowering time, plant height, number of nodes on the main stem, 100-seed weight, protein content, and oil content based on low-density molecular markers [16–18], fine-mapped an oil content locus on chromosome 15 [17] and a plant height locus on chromosome 13 [19], and cloned the key oil-related gene *GmSWEET39* [20]. Based on the *SojaCSSLP1*, we further developed an improved CSSL population *SojaCSSLP5* composed of 177 lines, 2,567,426 single nucleotide polymorphisms (SNPs), and we constructed 1366 SNP linkage disequilibrium block (SNPLDB) markers [21]. The development of these high-density molecular markers has laid a foundation for in-depth analysis of the molecular mechanisms of the protein content underlying the wild soybean.

The regulation mechanism of soybean protein content is closely related to biological processes such as lipid metabolism, carbon–nitrogen allocation, and seed development. A competitive relationship exists in the allocation of carbon sources between pathways synthesizing proteins and lipids, and this balance is determined by a regulatory network composed of multiple key genes, such as phosphoenolpyruvate carboxylase (*PEPC*) and stearoyl-ACP desaturase (*GmSACPD*). The former is a key enzyme linking sugar metabolism with amino acid synthesis, and its activity positively regulates protein content. The latter is a key enzyme in lipid synthesis, and its loss of function redirects carbon flux toward protein synthesis, leading to increased protein content and decreased oil content [22]. Transcriptional regulators include the transcription factors *POWR1* and the *GmDof* family. The former has been identified as a core regulatory switch, and its overexpression can significantly increase protein content, often at the expense of reduced oil content and seed size [23]. The *GmDof* family transcription factors play an important role in the protein–lipid balance by activating lipid synthesis genes and suppressing protein synthesis genes [24].

Nutrient transport proteins, such as the *GmSWEET* family genes (e.g., *GmSWEET10a/b*), are responsible for transporting photosynthetic products (e.g., sucrose) from the seed coat to the embryo. Their expression levels directly determine the supply of “raw materials” for seed filling and have pleiotropic effects on seed size, oil content, and protein content [25].

In summary, in the present study, we used the *SojaCSSLP5* population as material, through phenotypic evaluation in three environments and linkage analysis with high-density markers, aiming to accurately identify the chromosome segments in wild soybean that regulate high protein content and clarify their genetic effects and environmental stability and screen candidate genes for high protein content by integrating whole-genome resequencing data of both parents and transcriptome data from multiple tissues (leaves, flowers, seeds at different developmental stages, and pods). The results will provide materials for fine mapping and functional studies of related loci and offer a molecular basis for high-protein soybean breeding.

2. Materials and Methods

2.1. Plant Materials

The materials used in this study included cultivated soybean *NN1138-2*, wild soybean *N24852*, and the derived wild soybean CSSL population *SojaCSSLP5* containing 177 lines. *NN1138-2* belongs to maturity group V and is a key parent in Southern China soybean breeding, with a seed protein content of approximately 44.15%. *N24852* belongs to MG III, with a seed protein content of about 46.20%. The *SojaCSSLP5* was constructed using *NN1138-2* as the recipient and wild soybean *N24852* as the donor through continuous backcrossing and marker-assisted selection. The methodology for population construction and genotyping results was published in 2021 [21]. Whole-genome resequencing of *SojaCSSLP5* at $\sim 3.04\times$ was performed, producing 2,567,426 SNPs and 1366 SNPLDB markers. According to SNPLDB marker analysis, each line carries 1–24 wild segments, and the wild segments carried by the population cover 99.74% of the wild soybean genome [21].

2.2. Field Experiment Design

The *SojaCSSLP5*, along with its two parents, was planted in a complete randomized block design experiment with three replications, one row per plot, 1 m in length, 10 plants per row, and 0.5 m of row space at Jiangpu Experiment Station (Nanjing, China) in 2016 and 2017, and Dangtu Experiment Station (Maanshan, China) in 2018. The three environments were coded as 2016JP, 2017JP, and 2018DT, respectively.

2.3. Measurement and Phenotypic Analysis of Protein Content

Soybean seed protein content was determined using near-infrared spectroscopy (NIRS) with a FOSS InfratecTM 1255 NIR Grain Analyzer (Hillerod, Denmark). After harvest and drying, approximately 50 g of seeds from each plot were used. Each sample was measured three times, and the average value was taken as the phenotypic value of protein content. Descriptive statistics and analysis of variance were performed using the GLM procedure in SAS 9.4 (SAS/STAT software, SAS Institute Inc., Cary, NC, USA). Broad-sense heritability (h^2) was calculated as follows:

$$h^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_{GE}^2 / n + \sigma_e^2 / nr)$$

where σ_G^2 is the genetic variance, σ_{GE}^2 is the genotype-by-environment interaction variance, σ_e^2 is the error variance, n is the number of environments, and r is the number of replications per environment.

2.4. Identification of Wild Segments Associated with Protein Content

The stepwise regression based likelihood ratio test of additive QTL (RSTEP-LRT-ADD) model in the IciMapping 4.0 software was used to detect wild segments associated with protein content. This method is suitable for non-ideal CSSL populations. Its basic principle is using stepwise regression analysis to select the chromosome segments that have the greatest impact on the trait and the CSSLs carrying those segments, estimating the LOD value for each segment using the likelihood ratio, and mapping the trait-related segments based on the LOD threshold. QTL naming convention: q + trait abbreviation (capital letters) + chromosome number + QTL order number. The full QTL name is usually italicized, e.g., the third protein content QTL on chromosome 3 is denoted as *qPro3.3*.

2.5. Prediction of Candidate Genes for Seed Protein Content

To facilitate the prediction of candidate genes within the detected segments, our group previously conducted whole-genome resequencing and transcriptome sequencing of eight different tissues for both parents of the CSSL population, NN1138-2 and N24852. The tissues for transcriptome sequencing included: leaf, flower, and seeds at 14, 21, 28, and 35 days after flowering (14 seed, 21 seed, 28 seed, 35 seed). The sequencing depth was 5×, without biological or technical replicates. Candidate genes were predicted by comparing sequence variations (especially non-synonymous SNPs) and expression differences (≥ 2 -fold difference, particularly in late seed development stages) between the two parents, combined with functional annotations from Soybase (<http://www.soybase.org>).

3. Results

3.1. Performance of Protein Content Among the CSSL Population and Its Parents

The performance of protein content for the *SojaCSSLP5* and its parents across three environments (2016JP, 2017JP, 2018DT) is summarized in Table 1 and Figure 1. The protein content of NN1138-2 among the three environments was 45.48%, 43.98%, and 43.00%, respectively, with a mean value of 44.15%. The protein content of N24852 among the three environments was 46.97%, 45.88%, and 45.75%, respectively, with a mean value of 46.20%, indicating a significant difference between the two parents. Among the population, the 177 lines showed rich genetic variation in protein content averages over the three environments, with a mean of 45.00% and a range of 42.86–49.08%, indicating that the introduction of wild soybean chromosome segments caused widespread variation in the population. The distribution of protein content shows that most lines have reverted to the level of the recurrent parent NN1138-2, while only a few lines have protein content comparable to the donor parent N24852. This distribution characteristic is consistent with the features of a CSSL population. Furthermore, the error coefficient of variation for protein content across the three environments ranged between 0.03 and 0.04, indicating low experimental error and high data reliability. The heritability of protein content in the three environments was 0.75, 0.80, and 0.78, respectively, with an average heritability of 0.70, indicating that the variation in protein content in the population is mainly genetically controlled.

Table 1. The performance of *SojaCSSLP5* and its two parents in protein content.

Env.	Parents		SojaCSSLP5													Range (%)	Mean (%)	CV	h ²
	NN1138-2	N24852	Class Mid-Value (%)																
			41.4	42.2	43.0	43.8	44.6	45.4	46.2	47.0	47.8	48.6	49.4						
2016JP	45.48	46.97	0	1	11	30	43	35	16	11	5	1	1	42.22–49.27	45.08	0.03	0.75		
2017JP	43.98	45.88	0	3	13	52	43	29	18	4	2	0	1	42.13–49.15	44.60	0.03	0.80		
2018DT	43.00	45.75	5	20	34	37	29	9	11	2	4	2	1	41.40–49.15	44.00	0.04	0.78		
Mean	44.15	46.20	0	0	19	53	50	28	10	1	2	0	1	42.86–49.08	45.00	0.03	0.70		

2016JP, 2017JP, 2018DT and mean represent the environments of 2016 Jiangpu, 2017 Jiangpu, 2018 Dangtu, and average of the three environments, respectively.

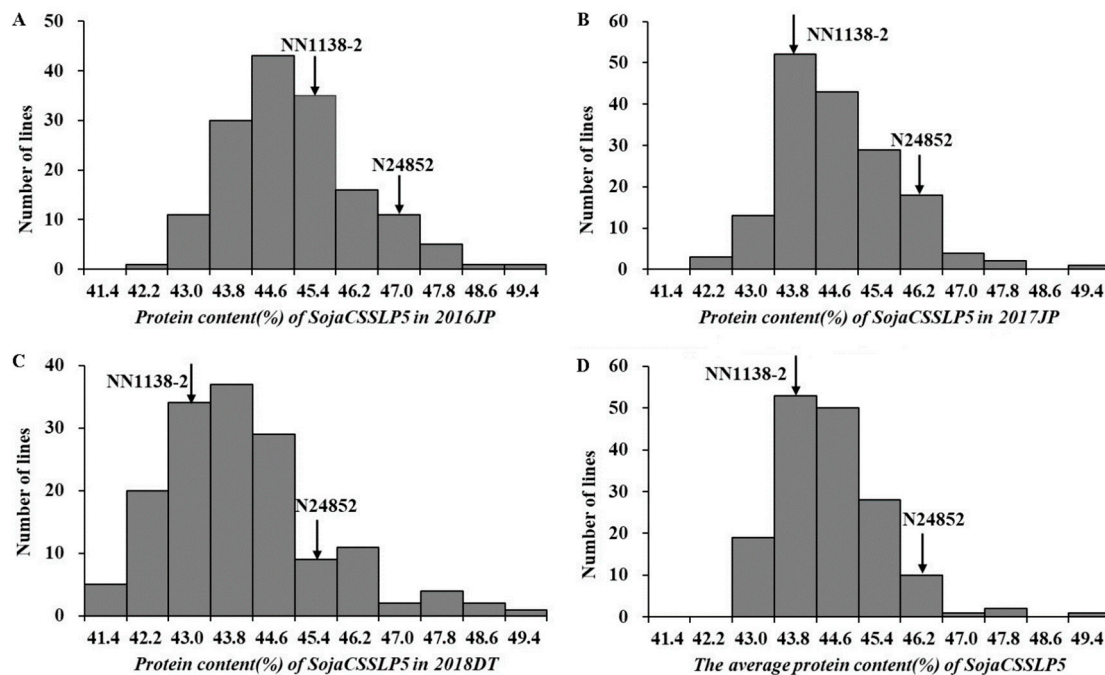


Figure 1. The frequency distributions of seed protein content in *SojaCSSLP5* along with its two parents. (A–D) indicate 2016JP, 2017JP, 2018DT and the average protein content of *SojaCSSLP5* respectively.

Joint analysis of variance for protein content across the three environments (Table 2) showed significant differences among environments, replicates within environments, lines, and line-by-environment interaction. These results indicate that although protein content is primarily genetically controlled, environmental conditions and genotype-by-environment interactions still significantly affect the phenotype. Therefore, using the mean phenotype across multiple environments for subsequent segment mapping can effectively reduce environmental interference and improve mapping accuracy.

Table 2. The joint ANOVA of protein content of the *SojaCSSLP5* under multiple environments.

Source of Variation	DF	MS	F Value	<i>p</i>
Environment (Env)	2	107.73	100.82	<0.0001
Repeat (Env)	6	17.01	15.92	<0.0001
Line	163	7.91	7.40	<0.0001
Line × Env	306	2.59	2.42	<0.0001
Error	837	1.07		
Total	1314			

Line is a fixed model; environment is a random model.

3.2. Identification of Wild Segments Associated with Protein Content

Based on the mean protein content of the *SojaCSSLP5* across three environments, six wild segments/QTL significantly associated with protein content were detected. They were distributed on five chromosomes: 3, 6, 9, 15, and 20, with two segments/QTL detected on chromosome 15. The significance (LOD value) of the segments/QTL ranged from 3.06 to 15.82. The segment/QTL *qPro20.1* on chromosome 20 was the most significant, with the LOD value as high as 15.82. The six detected segments/QTL collectively explained 57.42% of the phenotypic variation. The phenotypic variation explained (PVE) by individual loci ranged from 3.58% to 22.46%. Based on the PVE, the mapped segments can be classified into large-effect and small-effect segments. The segments/QTL on chromosomes 9, 15, and 20—*qPro9.1* (PVE 10.13%), *qPro15.1* (PVE 10.84%), and *qPro20.1* (PVE 22.46%)—had PVE exceeding 10% and are considered as core genetic loci regulating protein content in soybean. Among them, *qPro20.1* had a significantly higher contribution than the others, representing

the primary segment controlling this trait. The segments/QTL on chromosomes 3, 6, and 15—*qPro3.1* (PVE 3.58%), *qPro6.1* (PVE 5.74%), and *qPro15.2* (PVE 4.67%)—each had a PVE below 10% and are presumed to be minor genetic loci for protein content. Notably, the additive effect values for all six wild segments were positive, ranging from 0.42% to 1.09% (Table 3), indicating that these segments from the donor wild soybean N24852 can significantly increase the protein content of cultivated soybean. This is entirely consistent with the phenotypic characteristic that N24852 itself has a significantly higher protein content (46.20%) than the recipient parent NN1138-2 (44.15%), further validating the reliability of the mapping results. These segments are the key genetic loci of the high protein content trait in wild soybean.

Table 3. The QTL/segment related to protein content detected in the *SojaCSSLP5*.

QTL	Marker	Genome Region	Size of Region (KB)	LOD	PVE (%)	Add (%)	Reported QTL
<i>qPro3.1</i>	Gm03_LDB_36	40,609,452–41,028,039	418.59	3.06	3.58	0.42	Seed protein36-34 [26] Seed protein36-35 [26]
<i>qPro6.1</i>	Gm06_LDB_14	4,859,600–5,358,717	499.12	4.75	5.74	0.44	New
<i>qPro9.1</i>	Gm09_LDB_74	44,878,012–44,951,892	73.88	8.02	10.13	0.62	Seed protein36-28 [26] Seed protein36-29 [26]
<i>qPro15.1</i>	Gm15_LDB_12	5,335,613–6,085,762	750.15	8.49	10.84	0.92	Seed protein 30-3 [27]
<i>qPro15.2</i>	Gm15_LDB_31	11,936,865–12,087,185	150.32	3.93	4.67	0.66	Crude protein, R6 1-5 [28]
<i>qPro20.1</i>	Gm20_LDB_28	13,996,858–24,336,032	10,339.17	15.82	22.46	1.09	cqSeed protein-003 [29] Seed protein 1-3 [9] Seed protein 1-4 [9] Seed protein 3-12 [30] Seed protein 10-1 [10] Seed protein 11-1 [10] Seed protein 30-1 [27] Seed protein 31-1 [31] Seed protein 34-11 [32] Seed protein36-26 [26] Seed protein 37-8 [33]

3.3. QTL–Allele Matrix for Protein Content in the *SojaCSSLP5*

To systematically analyze the genetic distribution pattern of high-protein-related QTL alleles, a QTL–allele matrix for protein content was constructed for the *SojaCSSLP5* and its two parents (NN1138-2, N24852) based on the six mapped wild segments (*qPro3.1*, *qPro6.1*, *qPro9.1*, *qPro15.1*, *qPro15.2*, *qPro20.1*) (Figure 2). The matrix clearly defines the allele sources: black cells represent the positive alleles from the donor wild soybean N24852, and white cells represent the negative alleles from the recipient cultivated soybean NN1138-2, providing an intuitive presentation of the allelic combination pattern for each CSSL. Association analysis between the matrix and phenotypes revealed that the protein content level of the substitution lines showed a positive correlation trend with the number of positive wild segments they carried: lines with protein content significantly higher than the population mean (45.00%) generally carried two to three positive wild segments, whereas lines with protein content close to or lower than the recurrent parent (44.15%) mostly carried only zero to one positive wild segments and were predominantly composed of negative alleles, confirming the positive regulatory effect of the introgressed wild segments on protein content. Using the phenotypic values from the three environments (2016JP, 2017JP, 2018DT) for significance testing ($p < 0.05$), 41 lines with protein content significantly higher than the recurrent parent were screened out from the 177 CSSLs. Further allele tracing showed that among these 41 high-protein lines, 30 carried at least one of the high-protein-related wild segments mapped in this study. Among the mapping segments, the distribution frequency of the wild segment of *qPro20.1* (18 times) was the highest in

41 high-protein lines, followed by *qPro15.1* (11 times), indicating that these two segments are the core genetic factors driving the transgressive protein content in the lines. The remaining 123 low-protein lines (excluding black-seeded lines) cumulatively carried only 19 instances of protein-content-increasing wild segments, predominantly minor-effect segments (*qPro3.1*, *qPro6.1*) (14 times), with the frequency of large-effect segments being less than 4% (5 times). These results quantitatively demonstrate that the high-protein-related wild segments mapped in this study can explain 73.17% (30/41) of the phenotypic variation in the high-protein lines. This fully validates the genetic effectiveness of these mapping segments and provides clear genetic targets for subsequent marker-assisted selection (MAS) to pyramid positive segments and breed new high-protein germplasm.

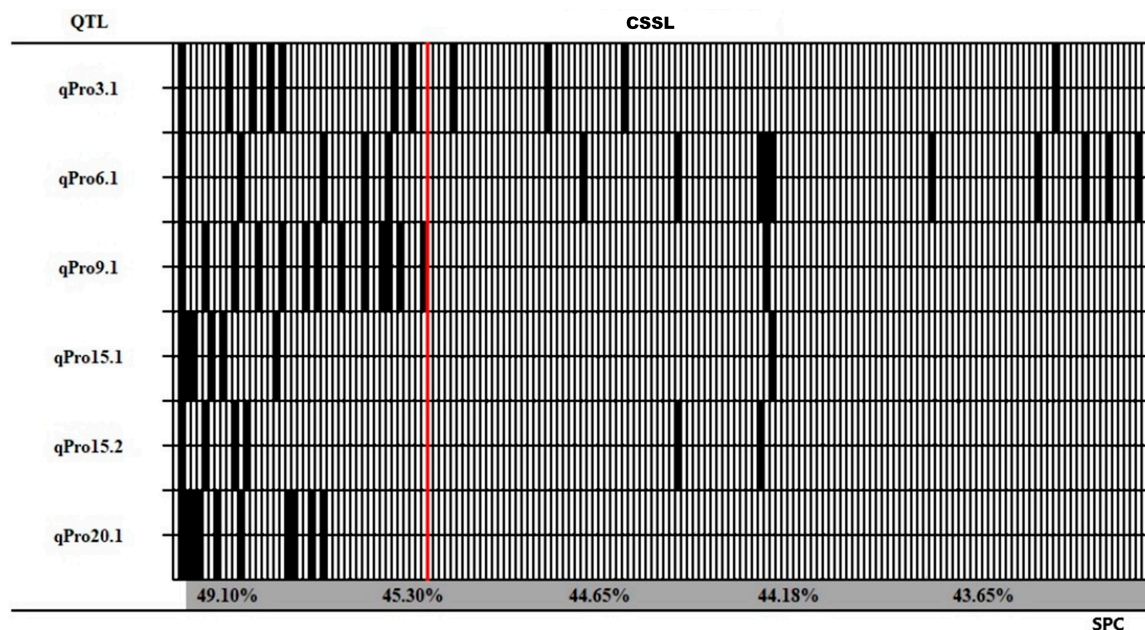


Figure 2. The QTL–allele matrix of seed protein content of the *SojaCSSLP5*. In CSSLs, the first column represents parent *NN1138-2*, the second column represents parent *N24852*, and the rest are all families of the population. A black cell represents a positive allele from the wild parent *N24852*, and a white cell represents a negative allele from the cultivated parent *NN1138-2*. The lines on the left of the solid red line in the figure have a significantly higher protein content than the parent *NN1138-2*.

3.4. Prediction of Candidate Genes for Soybean Protein Content

Based on the SoyBase database (<http://www.soybase.org>), candidate gene annotations were performed for the detected protein-content-related segments. A total of 362 candidate genes were contained within the six detected segments/QTL in this study. Using the parental transcriptome and genomic data obtained previously, candidate genes were predicted. First, genes with low expression (FPKM < 2.0) were filtered out based on RNA-seq data. Genes showing a ≥ 2 -fold expression difference between the parents, especially those with significant differences in expression at later stages of seed development, were selected. Second, by analyzing sequence differences between the parents using high-depth resequencing data, genes with SNP variations, particularly non-synonymous mutations, were identified as key candidate genes. Consequently, after screening within the six mapping segments, ten candidate genes for protein content were ultimately predicted (Table 4). These genes not only have sequence differences between the parents but also show significant differences in expression across various tissues, including leaf, flower, and seeds at 14, 21, 28, and 35 days after flowering. According to gene annotation and homologous gene analysis, some genes are involved in nucleic acid binding pathways (*Glyma.03G198500*), polynucleotide transferase pathways (*Glyma.03G200000*), and protein

binding pathways (*Glyma.06G066900*); some genes have protein transmembrane transport activity (*Glyma.09G223900*); others are related to serine-rich (*Glyma.06G065900*) and cysteine-rich proteins (*Glyma.09G224300*). The functions of these predicted key candidate genes still require further cloning and validation.

Table 4. Prediction of candidate genes related to the QTL of soybean protein content.

Gene	Gene Position	QTL	Annotation	Parents	SNP Haplotype	Expression Difference (FPKM)					
	Wm82.a2.v1					14 Seed	21 Seed	28 Seed	35 Seed	Leaf	Flower
<i>Glyma.03G198500</i>	Gm03:40774384–40776151	<i>qPro3.1</i>	Nucleic acid binding	NN1138-2 N24852	GCAC ATTG	13.34 13.89	12.94 7.04	4.50 14.87	8.91 58.44	15.30 20.17	9.52 8.33
<i>Glyma.03G200000</i>	Gm03:40887582–40889109	<i>qPro3.1</i>	Polynucleotidyl transferase/ribonuclease	NN1138-2 N24852	T C	6.93 5.11	4.16 1.31	1.31 3.04	1.68 10.76	4.91 5.11	9.05 14.85
<i>Glyma.06G065900</i>	Gm06:5028393–5034377	<i>qPro6.1</i>	Serine-rich protein-related	NN1138-2 N24852	CT AA	13.44 5.88	2.41 3.10	2.33 21.90	3.80 59.96	13.93 27.17	12.26 18.54
<i>Glyma.06G066900</i>	Gm06:5093927–5103072	<i>qPro6.1</i>	Protein binding	NN1138-2 N24852	CGAGTTCG AATTGATC	7.70 7.33	6.37 3.32	3.90 7.84	4.01 19.26	9.38 10.44	6.14 5.57
<i>Glyma.09G223900</i>	Gm09:44887793–44891771	<i>qPro9.1</i>	Protein transmembrane transport	NN1138-2 N24852	AAGT TGTA	6.42 4.37	6.33 2.18	3.75 6.65	3.49 26.63	3.32 3.52	2.82 2.61
<i>Glyma.09G224300</i>	Gm09:44916070–44919953	<i>qPro9.1</i>	Cysteine-rich protein-related	NN1138-2 N24852	CGAGCATT TACCGGAC	22.41 18.07	17.47 9.79	9.19 19.51	8.49 89.40	20.72 20.04	15.41 10.92
<i>Glyma.15G005900</i>	Gm15:503903–506167	<i>qPC15-1</i>	hydroxyproline-rich glycoprotein family protein	NN1138-2 N24852	GAAGTTCACG AGTTGATCGA	40.93 32.38	37.03 12.01	8.05 22.39	16.54 25.56	32.60 38.00	20.16 22.72
<i>Glyma.15G008800</i>	Gm15:700088–701291	<i>qPC15-1</i>	Embryo-specific protein 3	NN1138-2 N24852	CCAAGTAC ATTAGGT	99.22 89.02	88.49 55.65	19.82 30.93	22.64 1.15	67.61 45.05	97.39 163.15
<i>Glyma.15G145000</i>	Gm15:11930954–11932523	<i>qPro15.2</i>	-	NN1138-2 N24852	GA AG	15.08 31.51	21.69 19.94	10.93 9.96	17.49 0.25	41.46 25.42	12.79 15.84
<i>Glyma.20G085100</i>	Gm20:31774770–31779804	<i>qPro20.1</i>	Nucleoprotein containing a CCT domain	NN1138-2 N24852	- 321bp-InDel	0.59 2.23	1.13 2.26	0.57 1.45	1.31 0.19	0.27 1.96	2.75 0.64

FPKM represents fragments per kilobase per million.

4. Discussion

4.1. Genetic Basis of Protein Content in Wild Soybean and Comparison with Previous Studies

Wild soybeans have a higher protein content compared to cultivated soybeans, yet the genetic basis of their differences remains to be fully elucidated. This study, using a wild soybean CSSL population *SojaCSSLP5*, found that the genetic cause of the protein content difference between the wild parent *N24852* and the cultivated parent *NN1138-2* can be traced to six chromosomal regions. All wild segments had positive additive effects on protein content, collectively explaining 57.42% of the phenotypic variation. The remaining 12.58% (70–57.42%) of the phenotypic variation might be due to undetected minor QTLs. Among the six detected segments, the segments of *qPro9.1* (PVE 10.13%), *qPro15.1* (PVE 10.84%) and *qPro20.1* (PVE 22.46%), marked by *Gm09_LDB_74*, *Gm15_LDB_12*, and *Gm20_LDB_28*, respectively, had PVEs exceeding 10%. We infer that these three segments are large major-effect loci controlling the differentiation of protein content between wild and cultivated soybeans, while the remaining three segments are small major-effect loci.

Compared with previous studies (Table 4), five of the six wild segments mapped in this study contain QTLs of protein content previously reported [10,26–33]. The segment/QTL of *qPro20.1* mapped in this study encompasses 11 QTLs of protein content recorded in Soybase [10,26,27,29–33], derived from four cultivated/wild soybean combinations and six cultivated/cultivated soybean combinations. This indicates that the *qPro20.1* locus is differentiated not only between wild and cultivated soybeans but also among cultivated soybeans. Additionally, this study newly discovered a protein-content-increasing wild segment, *qPro6.1*, which was not detected in cultivated/cultivated soybean combinations, suggesting that this protein locus might be differentiated only between wild and cultivated soybeans.

The chromosome segment substitution line population *SojaCSSLP5* used in this study has been improved over four generations. Wang et al. [14] initially constructed a CSSL pop-

ulation, *SojaCSSLP1*, which was subsequently improved by Xiang et al. [16], He et al. [34], and Yang et al. [18], forming *SojaCSSLP2*, *SojaCSSLP3*, and *SojaCSSLP4*, respectively. By using *SojaCSSLP1*, four wild segments for protein content located on chromosomes 8, 14, 19, and 20, with the QTL on chromosome 20 explaining over 10% of the phenotypic variation [35]. By using *SojaCSSLP4* [18], nine protein content QTLs were detected on chromosomes 2, 4, 7, 8, 11, 13, 15, 17, and 20, respectively, with QTLs on chromosomes 2, 4, 15, 17, and 20 explaining over 10% of the phenotypic variation. Compared with the results of this study, *qPro20.1* overlaps with segments/QTLs detected in both *SojaCSSLP1* and *SojaCSSLP4*, and *qPro15.1* overlaps with a QTL detected in *SojaCSSLP4*. The other four QTLs were not detected in *SojaCSSLP1* and *SojaCSSLP4*. These results indicate that the soybean protein content locus *qPro20.1* detected in this study is a major locus controlling soybean protein content and should be a focus of future research. The other QTLs still need further verification by constructing secondary mapping populations from substitution lines and the recurrent parent NN1138-2.

4.2. Candidate Gene of Protein Content in Soybean

Soybean seed protein content is a complex quantitative trait controlled by multiple genes. These include amino acid uptake and synthesis, translation on the endoplasmic reticulum, processing and transport through the Golgi apparatus, and final deposition into protein storage vacuoles [36]. Consequently, genes regulating protein content are likely involved in diverse biological processes such as amino acid metabolism, protein translation and folding, membrane trafficking systems, and associated transcriptional regulation.

Based on the six wild segments associated with high protein content in this study, we systematically screened for candidate genes by integrating whole-genome resequencing data of the parents and transcriptome data from multiple tissues (including leaves, flowers, seeds at different developmental stages, and pods). Our screening strategy was primarily based on two key criteria: First, the presence of sequence variations in the coding region that could potentially affect protein function, particularly non-synonymous single nucleotide polymorphisms between the parents (NN1138-2 and N24852). Second, the significant differences in gene expression between the two parents (a fold-change ≥ 2), especially in key tissues like the middle and late stages of seed development. Through this strategy, we refined a list of 10 high-confidence candidate genes from the 362 genes located within the six QTL intervals (Table 4). Functional annotation and homologous gene analysis of these candidate genes revealed their involvement in various pathways potentially related to protein synthesis and accumulation. For instance, the gene *Glyma.03G200000*, located within the *qPro3.1* interval, encodes a protein with ribonuclease activity. Its homologous gene in rice has been reported to play a role in the transcriptional regulation of storage proteins [37], suggesting it might influence protein synthesis by modulating the stability of relevant mRNAs. The gene *Glyma.15G078300*, located within the *qPro15.1* interval, encodes a transcription factor containing an NAC domain. Studies have shown that NAC transcription factors in rapeseed are involved in responses to biotic and abiotic stresses [38], while NAC proteins in rice can directly activate the expression of genes related to starch and storage protein synthesis [39]. This indicates that *Glyma.15G078300* may act as an upstream regulator, directly activating the protein synthesis pathway in soybean seeds. The gene *Glyma.20G851000*, located within the *qPro20.1* interval, encodes a protein with ribonuclease activity. It has been identified as a core regulatory switch, and its overexpression can significantly increase protein content, often at the expense of reduced oil content and seed size [23]. The functions of other candidate genes involve nucleic acid binding (*Glyma.03G198500*), transmembrane protein transport (*Glyma.09G223900*), protein

kinase activity, and characteristics of serine/cysteine-rich proteins. These functions are closely associated with the synthesis, modification, transport, and accumulation of proteins.

It is noteworthy that among the 10 candidate genes predicted in this study, some genes and their homologs have been functionally validated to be associated with protein content in soybean, rice, and rapeseed, further strengthening their potential as key genes controlling soybean protein content. However, since the QTL intervals still contain multiple genes and functional redundancy may exist, the precise functions of these predicted candidate genes require further validation through molecular biological approaches such as developing secondary mapping populations, gene editing, and overexpression/knockdown experiments. The findings of this study provide precise targets and a solid theoretical foundation for subsequent gene cloning and functional studies, which will significantly advance molecular design breeding for high-protein soybean.

5. Conclusions

Annual wild soybean is characterized by its high protein content. In this study, six wild segments that can increase soybean protein content were detected based on a wild soybean chromosome segment substitution line population SojaCSSLP5. Among them, three were large-effect loci with phenotypic contribution rates beyond 10%. Among the wild segments detected in this study, five overlapped with protein-related loci previously reported in cultivated soybeans. The remaining segment, qPro6.1, represents a novel locus discovered in this study, suggesting that it exists only during soybean domestication. Furthermore, based on sequence variations and gene expression differences between the parents, ten candidate genes for protein content were predicted within the six detected segments, though their specific functions require further validation. These findings provide a foundation for mining and utilizing elite genes associated with a high protein content in wild soybeans.

Author Contributions: Conceptualization, J.G. and W.W.; methodology, N.L., M.C., W.L., W.H., C.L., J.H., F.L., L.S., G.X., J.G. and W.W.; software, N.L., W.L., W.H., C.L., J.H. and W.W.; validation, N.L., M.C., W.L., W.H., C.L., J.H., F.L., L.S., G.X., J.G. and W.W.; formal analysis, N.L., M.C., W.L., W.H., C.L., J.H., F.L., L.S., G.X., J.G. and W.W.; investigation, W.L., W.H., C.L. and W.W.; resources, J.G. and W.W.; data curation, N.L., M.C., W.L., W.H., C.L., J.G. and W.W.; writing—original draft preparation, N.L., M.C., W.L. and W.W.; writing—review and editing, N.L., W.L. and W.W.; visualization, W.L., W.H. and W.W.; supervision, J.G. and W.W.; project administration, J.G. and W.W.; funding acquisition, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City (2021JJLH0007), the National Key Research and Development Program of China (2024YFD201400), Zhongshan Biological Breeding Laboratory (ZSBBL-KY2023-03), the National Natural Science Foundation of China (32272147) and the MOA CARS-04 program (CARS-04).

Data Availability Statement: The corresponding phenotype and genotype data of the CSSLs can be accessed from the public website: <https://github.com/njau-sri/soybean-seed-protein-lining> (accessed on 14 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Foyer, C.H.; Lam, H.M.; Nguyen, H.T.; Siddique, K.H.M.; Varshney, R.K.; Colmer, T.D.; Cowling, W.; Bramley, H.; Mori, T.A.; Hodgson, J.M.; et al. Neglecting legumes has compromised human health and sustainable food production. *Nat. Plants* **2016**, *2*, 112. [CrossRef]
2. Liu, X.H.; Zhou, R.B.; Gai, J.Y. A comparative analysis of protein and fat content between wild and cultivated soybeans in China. *Soyb. Sci.* **2009**, *28*, 566–573.

3. Wang, J.L.; Meng, Q.X.; Yang, Q.K.; Zhao, S.W.; Wu, T.L. Effect of backcrossing on overcoming viny and lodging hablt of cultivated × wild and cultivated × semi-wild crosses. *Soyb. Sci.* **1986**, *5*, 181–187.
4. Hyten, D.L.; Pantalone, V.R.; Sams, C.E.; Saxton, A.M.; Landau-Ellis, D.; Stefaniak, T.R.; Schmidt, M.E. Seed quality QTL in a prominent soybean population. *Theor. Appl. Genet.* **2004**, *109*, 552–561. [[CrossRef](#)]
5. Wang, X.; Liao, H.; Yan, X. Study on analyzing soybean protein and oil content by neat-infrared spectroscopy. *Soyb. Sci.* **2005**, *24*, 199–201.
6. Bolon, Y.; Joseph, B.; Cannon, S.; Graham, M.; Diers, B.; Farmer, A.; May, G.; Muehlbauer, G.; Specht, J.; Tu, Z. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biol.* **2010**, *10*, 41. [[CrossRef](#)]
7. Fliege, C.; Ward, R.A.; Vogel, P.; Nguyen, H.; Quach, T.; Guo, M.; Viana, J.P.G.; Dos Santos, L.B.; Specht, J.E.; Clemente, T.E. Fine mapping and cloning of the major seed protein qtl on soybean chromosome 20. *Plant J.* **2022**, *10*, 1011–1025.
8. Marsh, J.I.; Hu, H.; Petereit, J.; Bayer, P.E.; Valliyodan, B.; Batley, J.; Nguyen, H.T.; Edwards, D. Haplotype mapping uncovers unexplored variation in wild and domesticated soybean at the major protein locus cqProt-003. *Theor. Appl. Genet.* **2022**, *153*, 1443–1455. [[CrossRef](#)] [[PubMed](#)]
9. Diers, B.W.; Keim, P.; Fehr, W.R.; Shoemaker, R.C. RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* **1992**, *83*, 608–612. [[CrossRef](#)]
10. Sebolt, A.M.; Shoemaker, R.C.; Diers, B.W. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Sci.* **2000**, *40*, 1438–1444. [[CrossRef](#)]
11. Eshed, Y.; Zamir, D. A genomic library of *Lycopersicon pennellii* in *L. esculentum*: A tool for fine mapping of genes. *Euphytica* **1994**, *79*, 175–179. [[CrossRef](#)]
12. Wan, X.Y.; Weng, J.F.; Zhai, H.Q.; Wang, J.K.; Lei, C.L.; Liu, X.L.; Guo, T.; Jiang, L.; Su, N.; Wan, J.M. Quantitative trait loci (QTL) analysis for rice grain width and fine mapping of an identified QTL allele *gw-5* in a recombination hotspot region on chromosome 5. *Genetics* **2008**, *179*, 2239–2252. [[CrossRef](#)]
13. Alpert, K.B.; Tanksley, S.D. High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: A major fruit weight quantitative trait locus in tomato. *Plant Biol.* **1996**, *93*, 15503–15507.
14. Wang, W.B.; He, Q.Y.; Yang, H.Y.; Xiang, S.H.; Zhao, T.J.; Gai, J.Y. Development of a chromosome segment substitution line population with wild soybean (*Glycine soja* Sieb. et Zucc.) as donor parent. *Euphytica* **2013**, *189*, 293–307. [[CrossRef](#)]
15. Jiang, H.W.; Li, C.D.; Li, R.C.; Li, Y.Y.; Yin, Y.B.; Ma, Z.Z.; Zeng, Q.L.; Zhang, W.B.; Liu, C.Y.; Chen, Q.S. Construction of wild soybean backcross introgression lines. *Chin. J. Oil Crop. Sci.* **2020**, *42*, 8–16.
16. Xiang, S.H.; Wang, W.B.; He, Q.Y.; Yang, H.Y.; Liu, C.; Xing, G.N.; Zhao, T.J.; Gai, J.Y. Identification of QTL/segments related to agronomic traits using CSSL population under multiple environments. *Sci. Agric. Sin.* **2015**, *48*, 10–22.
17. Yang, H.Y.; Wang, W.B.; He, Q.Y.; Xiang, S.H.; Tian, D.; Zhao, T.J.; Gai, J.Y. Identifying a wild allele conferring small seed size, high protein content and low oil content using chromosome segment substitution lines in soybean. *Theor. Appl. Genet.* **2019**, *132*, 2793–2807. [[CrossRef](#)]
18. Yang, H.Y.; Wang, W.B.; He, Q.Y.; Xiang, S.H.; Tian, D.; Zhao, T.J.; Gai, J.Y. Chromosome segment detection for seed size and shape traits using an improved population of wild soybean chromosome segment substitution lines. *Physiol. Mol. Biol. Plants* **2017**, *23*, 877–889. [[CrossRef](#)]
19. Zhang, X.L.; Wang, W.B.; Guo, N.; Zhang, Y.Y.; Bu, Y.P.; Zhao, J.M.; Xing, H. Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genom.* **2018**, *19*, 226. [[CrossRef](#)] [[PubMed](#)]
20. Miao, L.; Yang, S.; Zhang, K.; He, J.; Wu, C.; Ren, Y.; Gai, J.; Li, Y. Natural variation and selection in Gmsweet39 affect soybean seed oil content. *New Phytol.* **2020**, *225*, 1651–1666. [[CrossRef](#)]
21. Liu, C.; Chen, X.N.; Wang, W.B.; Hu, X.Y.; Han, W.; He, Q.Y.; Yang, H.Y.; Xiang, S.H.; Gai, J.Y. Identifying wild versus cultivated gene-alleles conferring seed coat color and days to flowering in soybean. *Int. J. Mol. Sci.* **2021**, *22*, 1559. [[CrossRef](#)]
22. Byfield, G.E.; Xue, H.; Upchurch, R.G. Two genes from soybean encoding soluble Δ9 stearoyl-ACP desaturases. *Crop Sci.* **2006**, *46*, 840–846. [[CrossRef](#)]
23. Goettel, W.; Zhang, H.; Li, Y.; Qiao, Z.; Jiang, H.; Hou, D.; An, Y.Q.C. POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nat. Commun.* **2022**, *13*, 3051. [[CrossRef](#)]
24. Sun, Q.; Xue, J.; Lin, L.; Liu, D.; Wu, J.; Jiang, J.; Wang, Y. Overexpression of soybean transcription factors GmDof4 and GmDof11 significantly increase the oleic acid content in seed of *Brassica napus* L. *Agronomy* **2018**, *8*, 222. [[CrossRef](#)]
25. Hu, Y.; Liu, Y.; Wei, J.-J.; Zhang, W.K.; Chen, S.Y.; Zhang, J.S. Regulation of seed traits in soybean. *Abiotech* **2023**, *4*, 372–385. [[CrossRef](#)] [[PubMed](#)]
26. Mao, T.T.; Jiang, Z.F.; Han, Y.P.; Teng, W.L.; Zhao, X.; Li, W.B. Identification of quantitative trait loci underlying seed protein and oil contents of soybean across multi-genetic backgrounds and environments. *Plant Breed.* **2013**, *132*, 630–641. [[CrossRef](#)]
27. Tajuddin, T.; Watanabe, S.; Yamanaka, N.; Harada, K. Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. *Breed Sci.* **2003**, *53*, 133–140. [[CrossRef](#)]

28. Asekova, S.; Kulkarni, K.P.; Kim, M.; Kim, J.H.; Song, J.T.; Shannon, J.G.; Lee, J.D. Novel quantitative trait loci for forage quality traits in a cross between *pi 483463* and 'hutcheson' in soybean. *Crop Sci.* **2016**, *56*, 2600–2611. [[CrossRef](#)]
29. Nichols, D.M.; Glover, K.D.; Carlson, S.R.; Specht, J.E.; Diers, B.W. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. *Crop Sci.* **2006**, *46*, 834–839. [[CrossRef](#)]
30. Brummer, E.C.; Graef, G.L.; Orf, J.; Wilcox, J.R.; Shoemaker, R.C. Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Sci.* **1997**, *37*, 370–378. [[CrossRef](#)]
31. Pandurangan, S.; Pajak, A.; Molnar, S.J.; Cober, E.R.; Dhaubhadel, S.; Hernandez-Sebastia, C.; Kaiser, W.M.; Nelson, R.L.; Huber, S.C.; Marsolais, F. Relationship between asparagine metabolism and protein concentration in soybean seed. *J. Exp. Bot.* **2012**, *63*, 3173–3184. [[CrossRef](#)]
32. Lu, W.G.; Wen, Z.X.; Li, H.C.; Yuan, D.H.; Li, J.Y.; Zhang, H.; Huang, Z.W.; Cui, S.Y.; Du, W.J. Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. *Theor. Appl. Genet.* **2013**, *126*, 425–433. [[CrossRef](#)] [[PubMed](#)]
33. Wang, X.Z.; Jiang, G.L.; Green, M.; Scott, R.A.; Song, Q.J.; Hyten, D.L.; Cregan, P.B. 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.* **2014**, *289*, 935–949. [[CrossRef](#)] [[PubMed](#)]
34. He, Q.Y.; Yang, H.Y.; Xiang, S.H.; Wang, W.B.; Xing, G.N.; Zhao, T.J.; Gai, J.Y. QTL mapping for the number of branches and pods using wild chromosome segment substitution lines in soybean [*Glycine max* (L.) Merr.]. *Plant Genet. Resour.* **2014**, *12*, S172–S177. [[CrossRef](#)]
35. Wang, W.B.; He, Q.Y.; Yang, H.Y.; Xiang, S.H.; Xing, G.N.; Zhao, T.J.; Gai, J.Y. Identification of QTL/segments related to seed-quality traits in *G. soja* using chromosome segment substitution lines. *Plant Genet. Resour.* **2014**, *12*, S65–S69. [[CrossRef](#)]
36. Derbyshire, E.; Wright, D.; Boulter, D. Legumin and vicilin, storage proteins of legume seeds. *Phytochemistry* **1976**, *15*, 3–24. [[CrossRef](#)]
37. Hayashi, S.; Wakasa, Y.; Ozawa, K.; Takaiwa, F. Characterization of *IRE1* ribonuclease-mediated mRNA decay in plants using transient expression analyses in rice protoplasts. *New Phytol.* **2016**, *210*, 1259–1268. [[CrossRef](#)] [[PubMed](#)]
38. Hegedus, D.; Yu, M.; Baldwin, D.; Gruber, M.; Sharpe, A.; Parkin, I.; Whitwill, S.; Lydiat, D. Molecular characterization of brassica napus NAC domain transcriptional activators induced in response to biotic and abiotic stress. *Plant Mol. Biol.* **2003**, *53*, 383–397. [[CrossRef](#)]
39. Wang, J.; Chen, Z.C.; Zhang, Q.; Meng, S.S.; Wei, C.X. The NAC transcription factors *OsNAC20* and *OsNAC26* regulate starch and storage protein synthesis. *Plant Physiol.* **2020**, *184*, 1775–1791. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.