International Journal of
Molecular Sciences

MDPI

Article

Quantitative Trait Loci (QTL) Analysis of Seed Protein and Oil Content in Wild Soybean (*Glycine soja*)

Woon Ji Kim [1], Byeong Hee Kang [1,2], Chang Yeok Moon [1,2], Sehee Kang [1,2], Seoyoung Shin [1], Sreeparna Chowdhury [1], Man-Soo Choi [3], Soo-Kwon Park [3], Jung-Kyung Moon [3] and Bo-Keun Ha [1,2,*]

[1] Department of Applied Plant Science, Chonnam National University, Gwangju 61186, Republic of Korea
[2] BK21 FOUR Center for IT-Bio Convergence System Agriculture, Chonnam National University, Gwangju 61186, Republic of Korea
[3] National Institute of Crop Science, Rural Development Administration (RDA), Wanju 55365, Republic of Korea
* Correspondence: bkha@jnu.ac.kr; Tel.: +82-62-530-2055

Abstract: Soybean seeds consist of approximately 40% protein and 20% oil, making them one of the world's most important cultivated legumes. However, the levels of these compounds are negatively correlated with each other and regulated by quantitative trait loci (QTL) that are controlled by several genes. In this study, a total of 190 F_2 and 90 BC_1F_2 plants derived from a cross of Daepung (*Glycine max*) with GWS-1887 (*G. soja*, a source of high protein), were used for the QTL analysis of protein and oil content. In the $F_{2:3}$ populations, the average protein and oil content was 45.52% and 11.59%, respectively. A QTL associated with protein levels was detected at Gm20_29512680 on chr. 20 with a likelihood of odds (LOD) of 9.57 and an R^2 of 17.2%. A QTL associated with oil levels was also detected at Gm15_3621773 on chr. 15 (LOD: 5.80; R^2: 12.2%). In the $BC_1F_{2:3}$ populations, the average protein and oil content was 44.25% and 12.14%, respectively. A QTL associated with both protein and oil content was detected at Gm20_27578013 on chr. 20 (LOD: 3.77 and 3.06; R^2 15.8% and 10.7%, respectively). The crossover to the protein content of $BC_1F_{3:4}$ population was identified by SNP marker Gm20_32603292. Based on these results, two genes, *Glyma.20g088000* (S-adenosyl-L-methionine-dependent methyltransferases) and *Glyma.20g088400* (oxidoreductase, 2-oxoglutarate-Fe(II) oxygenase family protein), in which the amino acid sequence had changed and a stop codon was generated due to an InDel in the exon region, were identified.

Keywords: wild soybean; protein; oil; quantitative trait loci; indels

check for updates

Citation: Kim, W.J.; Kang, B.H.; Moon, C.Y.; Kang, S.; Shin, S.; Chowdhury, S.; Choi, M.-S.; Park, S.-K.; Moon, J.-K.; Ha, B.-K. Quantitative Trait Loci (QTL) Analysis of Seed Protein and Oil Content in Wild Soybean (*Glycine soja*). *Int. J. Mol. Sci.* **2023**, *24*, 4077. https://doi.org/10.3390/ijms24044077

Academic Editor: Jianhua Zhang

Received: 30 January 2023
Revised: 10 February 2023
Accepted: 13 February 2023
Published: 17 February 2023

Copyright: © 2023 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* L.) is an important crop worldwide that represents a major source of protein and vegetable oil for the human diet and animal feed. As of 2021, soybean was the largest source of protein meal in the world, at 243.6 t, and the second-largest source of vegetable oil (58.7 t) after palm oil (http://soystats.com/, accessed on 12 February 2023). In Asian countries, soybean seeds are used to produce a number of food products, including soymilk, tofu, soybean paste, natto, and soy sauce. In the West, soybean is typically used for soybean meal and seed oil. Soybean seeds generally consist of 40% protein and 20% oil [1], and these traits are negatively correlated with each other [2,3]. For this reason, it is very difficult to improve both traits simultaneously. In addition, because there is a negative correlation between seed yield and protein content [4], high-protein varieties need to be developed with care. In addition to being easily influenced by environmental factors, the protein and oil content of soybean seeds is regulated by the polygenes and quantitatively inherited [5,6]. These polygenes can be divided into major genes, which are less influenced by the environment and that have a significant influence on these levels; and minor genes, which have a weaker influence.

Wild soybean (*G. soja* Sieb. and Zucc.), the ancestor of cultivated soybean, has high genetic diversity and is thus valuable as a breeding material for soybean breeding programs [7,8]. Various studies have used wild soybean to improve biological stress resistance, abiotic stress tolerance, nutrition, and yields [9]. The average protein content of wild soybean is reported to be higher than that of cultivated soybean, although this may be due to correlations with the yield or oil content [9]. In the study by Chen and Nelson (2004), the protein content of wild and cultivated soybean lines was about 47% and 40%, respectively, while the oil content was 15% and 11%, respectively [10].

After the publication of the soybean genome for the first time by Schmutz et al. (2010) [11], Ha et al. (2012) [12] advanced genomic research further with the integration of physical maps for *G. max* and *G. soja*. QTL mapping uses F_2, backcross (BC), and recombinant inbred line (RIL) populations derived from bi-parental crosses. In many soybean populations, the QTLs for proteins and oils have been mapped to genomic regions on chromosomes 15 and 20 [13–17]. Major QTLs for protein and oil content were identified by Diers et al. (1992) using RFLP markers for the F2 population through the crossbreeding of *G. max* (A81-356022) and *G. soja* (PI 468916) [18]. The QTL located on chromosome 15 has been fine-mapped at an interval of 535 kb between simple sequence repeat (SSR) markers (Kim et al. 2016) [19], while the candidate gene for the QTL located on chromosome 20 has been cloned as *Glyma.20G85100* encoding the CCT domain [20].

To date, 255 and 322 protein and oil content-related QTLs, respectively, have been identified using bi-parental populations (https://www.soybase.org/, accessed on 12 February 2023). However, these QTLs may include multiple duplicate detections, so the Soybean Genetics Committee has emphasized the importance of experimentally confirming QTLs and adding "*cq*" in front of the original QTL name to indicate that it has been confirmed [20,21]. In total, 16 QTLs each have been confirmed for protein and oil content (https://www.soybase.org/), and these are distributed across 11 chromosomes, including chromosomes 15 and 20 [18–20]. Of these, the only QTLs derived from wild soybean are *cqpro*-003 and *cqoil*-004 [22]. Therefore, the purpose of this study was to discover new genes for protein and oil content from wild soybean using two progeny types derived from high-protein wild soybean lines.

2. Results

2.1. Phenotypic Variation in the Protein and Oil Content

In the present study, the seed oil and protein content were measured using seeds harvested from F_3 and BC_1F_3 progeny lines in 2019 and 2020, respectively. The protein content of Daepung and GWS-1887, the parents of the $F_{2:3}$ line, in 2019 was 37.10% and 50.37%, respectively, while the oil content was 19.81% and 5.83%, respectively. In the 190 $F_{2:3}$ plants, the protein content ranged from 40.08% to 50.96% with a mean of 45.52%, and the oil content ranged from 7.84% to 16.61% with a mean of 11.59% (Table 1).

Table 1. Seed protein and oil content of the $F_{2:3}$ mapping population and its parental lines Daepung and GWS-1887 in 2019.

	Daepung	GWS-1887	Daepung × GWS-1887 $F_{2:3}$				
			Min	Max	Average	Skew	Kurt
Protein (%)	37.10	50.37	40.08	50.96	45.52	0.08	0.04
Oil (%)	19.81	5.83	7.84	16.61	11.59	0.15	1.04

Min, minimum; Max, maximum; Skew, skewness; Kur, Kurtosis.

The protein content of Daepung and GWS-1887, the parents of the $BC_1F_{2:3}$ line, in 2020 was 40.05% and 49.28%, respectively, and the oil content was 16.53% and 5.34%, respectively. In the BC_1F_2 population, the protein content ranged from 31.50% to 49.54% with a mean of 44.25% and the oil content ranged from 7.73% to 14.84% with a mean of 12.14% (Table 2).

Table 2. Seed protein and oil content of the $BC_1F_{2:3}$ mapping population and its parental lines Daepung and GWS-1887 in 2020.

	Daepung	GWS-1887	Daepung × GWS-1887 $BC_1F_{2:3}$				
			Min	Max	Average	Skew	Kurt
Protein (%)	40.05	49.28	31.50	49.54	44.25	−1.32	4.40
Oil (%)	16.53	5.34	7.73	14.84	12.14	−0.47	0.55

Min, minimum; Max, maximum; Skew, skewness; Kur, Kurtosis.

The phenotypic variation of the F_2 and BC_1F_2 populations followed a normal distribution (Figures 1 and 2, respectively).

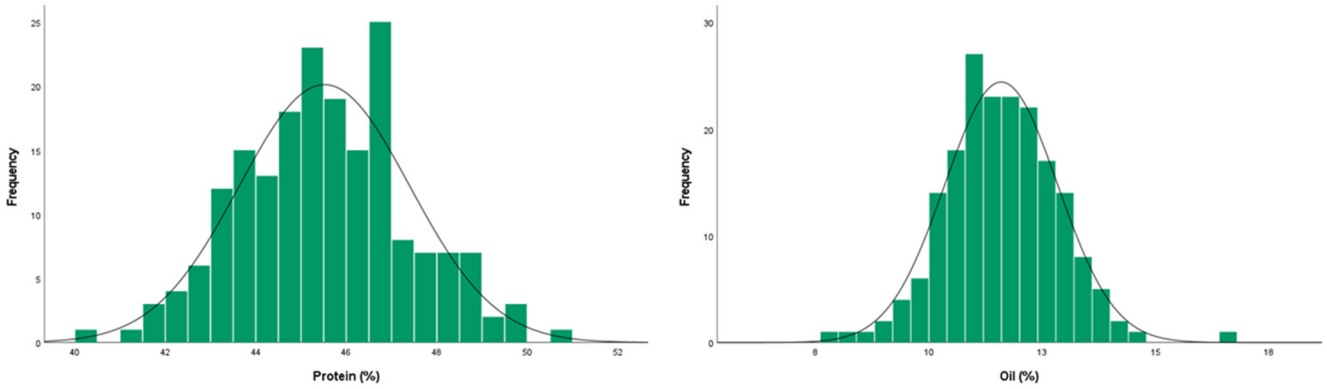


Figure 1. Distribution of the seed protein and oil content in the $F_{2:3}$ mapping population derived from a cross between Daepung and GWS-1887 in 2019.

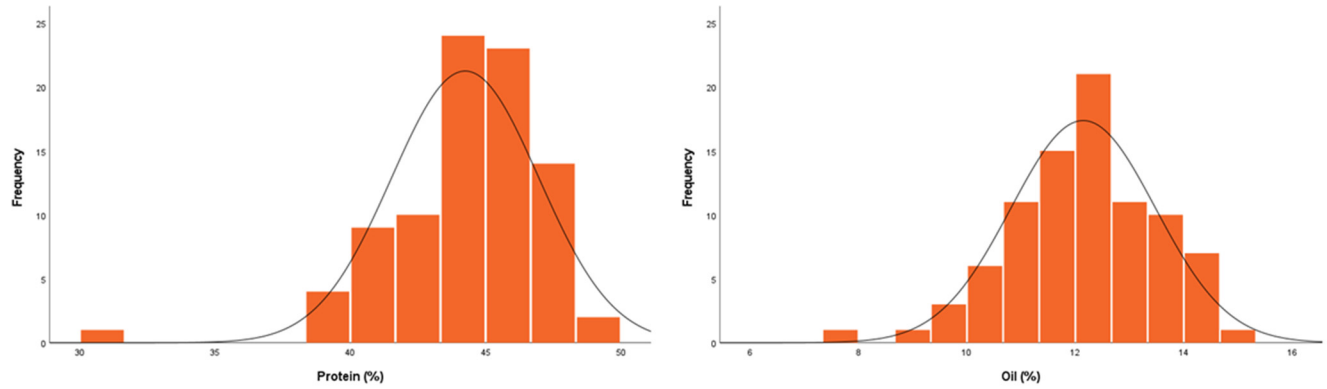


Figure 2. Distribution of the seed protein and oil content in the $BC_1F_{2:3}$ mapping population derived from a cross between Daepung and GWS-1887 in 2020.

2.2. Linkage Maps and QTL Analysis

Linkage maps for the F_2 and BC_1F_2 populations were constructed using polymorphic SNP markers acquired from SoySNP6K Illumina BeadChips (Figures S1 and S2, respectively). In the F_2 population, 2592 polymorphism markers were used, with an average chromosome length of 95 cM and an average of 130 markers located across each of the 20 chromosomes (Table 3).

Table 3. Summary of the genetic linkage map for the F_2 mapping population derived from a cross between Daepung and GWS-1887.

Chr.	Marker No.	Length (cM)	Average Spacing (cM)	Chr.	Marker No.	Length (cM)	Average Spacing (cM)
1	105	129.0	1.2	11	142	154.2	1.1
2	110	168.4	1.5	12	106	134.3	1.3
3	130	146.7	1.1	13	165	161.1	1.0
4	92	131.1	1.4	14	138	126.6	0.9
5	92	126.8	1.4	15	134	141.5	1.1
6	140	173.7	1.2	16	110	113.7	1.0
7	158	141.5	0.9	17	131	138.6	1.1
8	182	181.6	1.0	18	140	136.9	1.0
9	113	133.6	1.2	19	138	151.0	1.1
10	153	158.2	1.0	20	116	148.9	1.3
				Total	2595	2897.4	1.1

In the BC_1F_2 population, 1063 polymorphism markers were used, with an average chromosome length of 60 cM (except chromosome 12, which did not show polymorphism) and an average of 130 markers located on each of the 19 chromosomes (Table 4).

Table 4. Summary of the genetic linkage map for the BC_1F_2 mapping population derived from a cross between Daepung and GWS-1887.

Chr.	Marker No.	Length (cM)	Average Spacing (cM)	Chr.	Marker No.	Length (cM)	Average Spacing (cM)
1	38	51.9	1.4	11	49	48.7	1.0
2	54	48.8	0.9	12	-	-	-
3	73	70.1	1.0	13	112	125.1	1.1
4	59	70.0	1.2	14	40	36.5	0.9
5	78	97.2	1.2	15	90	104.8	1.2
6	11	10.8	1.0	16	43	45.6	1.1
7	23	20.7	0.9	17	30	33.0	1.1
8	57	57.0	1.0	18	81	70.6	0.9
9	44	44.2	1.0	19	31	30.0	1.0
10	67	68.6	1.0	20	83	101.8	1.2
				Total	1063	1135.4	1.1

The average genetic interval for both the F_2 and BC_1F_2 populations was 1.1 cM. QTLs for the protein and oil content in both populations were identified using MQM mapping analysis. In the $F_{2:3}$ population, QTLs for protein and oil content were identified on chromosomes 20 and 15, respectively (Figure 3), and these QTLs accounted for 17.2% and 12.2% of the phenotypic variation, respectively, with additive effects of the alleles on these traits of −1.10 and 0.59 (Table 5).

Table 5. Effects of the SNP markers associated with seed protein and oil content in the $F_{2:3}$ population.

SNP	Chr.	LOD	Homo AA	Homo BB	R^2 (%)
Protein					
Gm20_29512680	20	9.57	44.53	46.68	17.2
Oil					
Gm15_3621773	15	5.80	12.05	10.88	12.2

SNP, single nucleotide polymorphism; Chr, chromosome; LOD, likelihood of odds; Homo AA, Daepung allele; Homo BB, GWS-1887 allele.

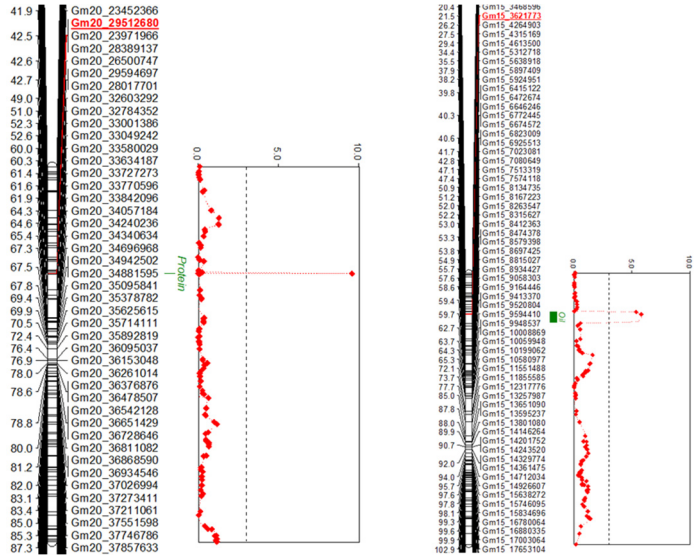


Figure 3. Likelihood of odds (LOD) plot for the seed protein and oil content using an LOD threshold of 3.0 (the vertical dotted line). These QTLs were mapped in the $F_{2:3}$ population derived from a cross between Daepung and GWS-1887.

On the other hand, in the $BC_1F_{2:3}$ population, QTLs for protein and oil content were both identified on chromosome 20 (Figure 4), accounting for 15.8% and 10.7% of the phenotypic variation, respectively, with additive effects of the alleles on these traits of −1.49 and 0.62 (Table 6).

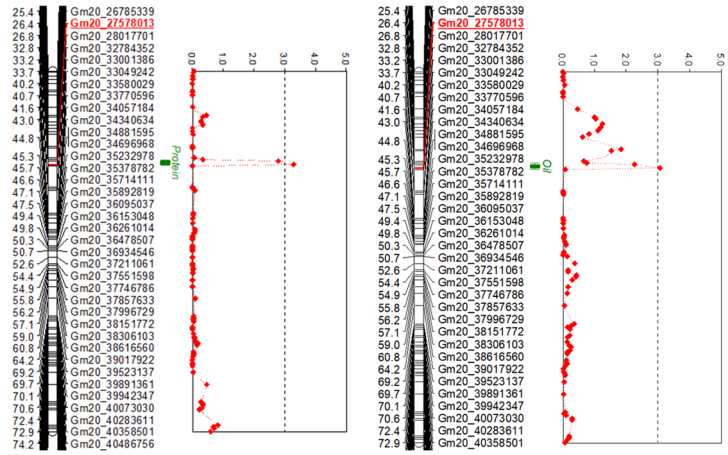


Figure 4. Likelihood of odds (LOD) plot for the seed protein and oil content using an LOD threshold of 3.0 (the vertical dotted line). These QTLs were mapped in the $BC_1F_{2:3}$ population derived from a cross between Daepung and GWS-1887.

Table 6. Effects of the SNP markers associated with seed protein and oil content in the $BC_1F_{2:3}$ population.

SNP	Chr.	LOD	Homo AA	Homo BB	R^2 (%)
Protein					
Gm20_27578013	20	3.29	42.66	45.63	15.8
Oil					
Gm20_27578013	20	3.06	12.70	11.47	10.7

SNP, single nucleotide polymorphism; Chr, chromosome; LOD, likelihood of odds; Homo AA, Daepung allele; Homo BB, GWS-1887 allele.

2.3. Crossover Detection

From the BC_1F_3 population, two lines with a heterozygote at a position expected to represent a high-protein-promoting gene on chromosome 20 were selected and advanced to produce a $BC_1F_{3:4}$ generation, followed by genotyping and protein content measurement. In Table 7, crossover occurred at 33,049,242 bp, and the individual with the genotype of the parent Daepung had average protein levels of 44.10 g, while individuals with the genotype of the parent GWS-1887 had a protein level of 47.58 g. In addition, crossover occurred at 32,603,292 bp, and individuals with the genotype of the parent Daepung had average protein levels of 45.25 g, and the individual with the genotype of the parent GWS-1887 had a protein level of 48.54 g. Given the genotypes of the two lines, it was predicted that the range of genes related to high protein levels is present at least downstream of position 32,603,292 bp.

Table 7. Markers and genotypes of BC_1F_4, two high-protein line selected from BC_1F_3 and advanced in generations.

Maker Name	BC_1F_4 Population No.																		
	141-1	141-3	141-5	141-6	141-7	141-8	141-10	141-12	141-14	145-1	145-3	145-4	145-5	145-7	145-9	145-10	145-11	145-14	145-15
BARC_1.01_Gm20_26500747	A †	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_26785339	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_27578013	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_28017701	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_28070832	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_28389137	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_29512680	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_29594697	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_32603292	A	A	A	A	B	A	A	A	A	A	B	A	A	A	B	A	B	B	B
BARC_1.01_Gm20_32784352	A	A	A	A	B	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33001386	A	A	A	A	B	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33049242	A	A	A	A	B	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33580029	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33634187	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33727273	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33770596	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_33842096	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
BARC_1.01_Gm20_34057184	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Protein content (%)	44.40	44.81	45.47	43.90	47.58	44.32	43.38	43.60	42.94	45.87	50.17	45.55	45.66	43.92	47.43	45.24	48.60	48.04	48.48

† 'A' genotype designates that the selected BC_1F_4–derived line was homogeneous for the allele from Daepung, 'B' designates that the line was homogeneous for the allele from GWS-1887.

2.4. Genome Re-Sequencing

Genome-resequencing analysis was conducted on wild soybean GWS-1887, which has a high protein content. The total number of sequencing reads was about 260 million with a sequencing depth of 38.6× and a total size of about 39 billion bp, while the coverage for the reference genome was 95.3%. Compared with the Williams 82 reference genome, approximately 4.7 million SNPs and 0.9 million InDels were identified in GWS-1887 (Table 8).

Table 8. Statistics for the high-quality reads from GWS-1887 mapped to the reference soybean genome.

	GWS-1887
Total reads	260,761,086
Total size (bp)	39,374,923,986
Mapped reads (%)	98.8
Genome coverage (%)	95.3
Sequencing depth	38.6×
Number of SNPs	4,750,431
Number of InDels	897,005

A total of 21 protein-related candidate genes with InDels were detected between Gm20_27578013, which is a molecular marker identified as a result of QTL mapping in the $BC_1F_{2:3}$ population, and Gm20_32603292, which was identified as the crossover site in $BC_1F_{3:4}$ (Table 9).

Table 9. Candidate genes for seed protein content and InDels between the reference genome and GWS-1887.

Gene	Gene Position (bp)	Annotation	Position of Genetic Variation (bp)	GWS-1887
Glyma.20G082450	31,080,736..31,082,822	Ammonium transporter in embryo development	31082569	small indel
Glyma.20G084000	31,486,240..31,488,766	Small nuclear ribonucleoprotein F	31486478	missing allele
			31486484	missing allele
			31486498	missing allele
			31486737	ref
			31487037	ref
			31487188	ref
			31487573	ref
			31488329	ref
			31488590	ref
Glyma.20G084051	31,490,248..31,494,204	Protein FAR1-RELATED SEQUENCE 5-like	31490350	small indel
			31493215	small indel
Glyma.20G084100	31,538,100..31,541,615	Tetratricopeptide repeat (TPR)-like superfamily protein	31541401	small indel
			31541587	small indel
Glyma.20G084500	31,616,022..31,624,831	Pre-mRNA-processing factor 17-like isoform X1	31624215	small indel
Glyma.20G085100	31,724,592..31,729,626	CCT motif family protein	31727019	ref
Glyma.20G085250	31,749,953..31,750,702		31750696	small indel
Glyma.20G085300	31,772,592..31,776,697	Uncharacterized protein LOC100814338 isoform X3	31772869	small indel
			31775624	small indel
			31775992	large indel
Glyma.20G085400	31,778,024..31,782,508	60S ribosomal L23-like protein	31778202	small indel
			31778572	small indel
			31779221	small indel
			31779277	small indel
			31780211	small indel
			31780626	small indel
			31781580	small indel
			31782506	small indel
Glyma.20G085450	31,783,157..31,786,246	Uncharacterized protein LOC102669815	31783447	large indel
			31784986	small indel
Glyma.20G085500	31,799,904..31,802,040	HAD superfamily	31801435	small indel
Glyma.20G085700	31,931,514..31,933,982	Unknown protein	31931542	small indel
			31932782	small indel
			31933822	small indel
Glyma.20G085800	31,936,253..31,941,630	Eukaryotic translation initiation factor 4E	31940258	small indel
			31940854	large indel
Glyma.20G086100	31,963,336..31,965,943		31965594	ref
Glyma.20G086800	32,280,129..32,280,768	PA domain	32280386	small indel
Glyma.20g086900	32,344,475..32,346,588	Aldehyde dehydrogenase	32345329	small indel
			32345819	ref
			32346510	small indel
Glyma.20g087000	32,482,132..32,486,538	Signal transduction histidine kinase	32482188	small indel
			32482242	small indel
			32482250	small indel
			32482352	small indel
			32482410	small indel

Table 9. *Cont.*

Gene	Gene Position (bp)	Annotation	Position of Genetic Variation (bp)	GWS-1887
			32482424	small indel
			32482464	small indel
			32482474	small indel
			32482480	small indel
			32482487	small indel
			32482549	small indel
			32482603	small indel
			32482697	small indel
			32482745	small indel
			32482761	small indel
			32482765	small indel
			32482811	small indel
			32482815	small indel
			32482850	small indel
			32482968	small indel
			32482972	small indel
			32482985	small indel
			32483011	small indel
			32483034	small indel
			32483036	small indel
			32483056	small indel
			32483091	small indel
			32483140	small indel
			32483171	small indel
			32483226	small indel
			32483233	small indel
			32483237	small indel
			32483278	small indel
			32483319	small indel
			32483325	small indel
			32483396	small indel
			32483411	large indel
			32483458	small indel
			32483552	small indel
			32485633	small indel
			32486435	small indel
Glyma.20G087600	32,693,094..32,698,696	Target SNARE coiled-coil domain protein	32694336	small indel
			32696653	small indel
Glyma.20G088000	32,881,065..32,886,678	S-adenosyl-L-methionine-dependent methyltransferases	32881614	small indel
			32882108	small indel
			32882111	large indel
			32882633	small indel
			32882909	small indel
			32883949	ref
			32883973	small indel
			32884174	small indel
			32884324	small indel
			32884601	small indel
			32885578	ref
			32885582	small indel
			32885642	ref
			32885647	ref

Table 9. *Cont.*

Gene	Gene Position (bp)	Annotation	Position of Genetic Variation (bp)	GWS-1887
			32885839	ref
			32885841	small indel
			32885843	small indel
			32885941	small indel
			32886302	large indel
			32886312	small indel
			32886569	large indel
Glyma.20G088100	32,887,478..32,888,573	(S)-coclaurine N-methyltransferase-like	32887547	small indel
			32887824	ref
Glyma.20G088400	32,909,361..32,910,905	Oxidoreductase	32910800	small indel

Of these, large InDels were identified in five genes (*Glyma.20G085300*, *Glyma.20G085450*, *Glyma.20G085800*, *Glyma.20g087000*, and *Glyma.20G088000*), small InDels were most common in 40 loci in *Glyma.20G085300*, and three large InDels were present in *Glyma.20G088000*. In particular, InDels occurred in the exon region of genes *Glyma.20g088000* and *Glyma.20g088400*, and the InDels of the two genes generated stop codons with amino acid frameshifts (Figure 5).

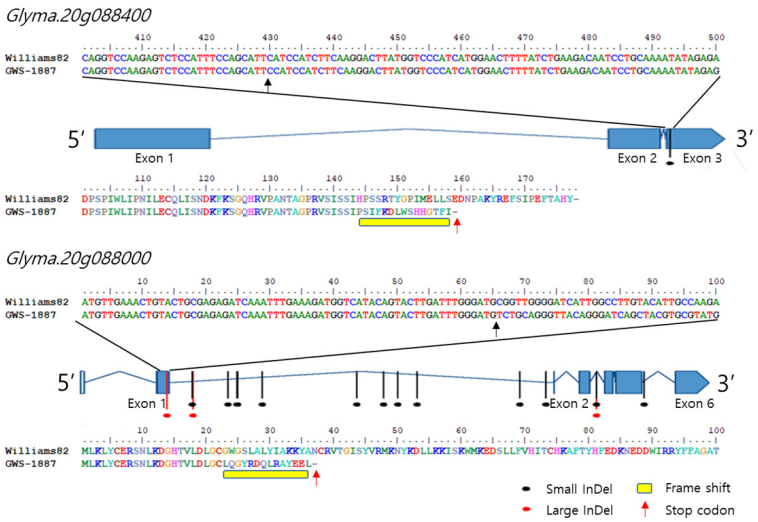


Figure 5. InDels in the *Glyma.20g088400* and *Glyma.20g088000* genes from GWS-1887.

Of these, the stop codon in *Glyma.20g088000* is expected to greatly simplify the structure of the protein (Figure 6).

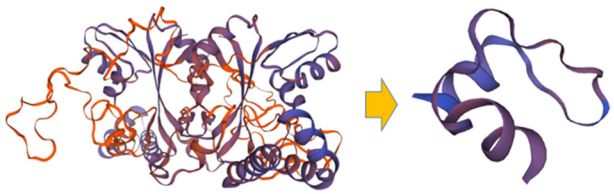


Figure 6. 3D prediction of the structure of the *Glyma.20g088000* protein from GWS-1887.

3. Discussion

QTL analysis of the protein and oil content in soybean has been well-studied in previous research. The present study searched for QTLs related to high protein content using two F_2 and BC_1F_2 populations derived from a cross between cultivated soybean

variety Daepung and wild soybean variety GWS-1887. The protein content of cultivated soybean is known to be about 40% [1], whereas wild soybean GWS-1887 has protein levels close to 50% [9,10,23]. This suggests that wild soybean GWAS-1887 may be useful for QTL analysis in terms of mapping the genetic regions associated with high protein content in soybean. However, the crossbreeding between *G. max* and *G. soja* may lead to linkage drag and consequent negative introgression such as reduced yields [9]. Daepung exhibited a higher annual variation in its protein content (37.10% in 2019 and 40.05% in 2020) than did GWS-1887 (50.37% in 2019 and 49.28% in 2020) (Tables 1 and 2). Several studies have reported that a lack of soil moisture reduces the protein content of soybean [24–26]. The rainfall in the experimental area in August 2019 during the soybean development period was lower than normal, while the rainfall in August 2020 was above average (http://www.kma.go.kr/, accessed on 12 February 2023). Therefore, it appears that the protein content of wild soybean has a higher environmental stability than cultivated soybean.

Previous QTL analysis of the protein and oil content in soybean seeds has been conducted with various populations, with the identified QTLs distributed across 20 chromosomes (https://www.soybase.org/). Of these, major QTLs for protein and oil content are present on chromosomes 15 and 20 [18], and several researchers have attempted to narrow down their precise location [13,19–22,27]. In the present study, the QTLs related to protein and oil content in the $F_{2:3}$ population were identified as Gm20_29512680 and Gm15_3621773, respectively, whereas in the $BC_1F_{2:3}$ population, marker Gm20_27578013 was identified for both protein and oil. Kim et al. (2016) reported the fine-mapping of a backcross line of Williams 82 and PI 407788A with 96 BARCSOYSSR markers and found that the QTL related to the protein and oil content on chromosome 15 was present in a 535 kb region from the physical position 3.59 Mbp to 4.12 Mbp [19]. These results are consistent with the SNP marker Gm15_2621773 at the physical position 3.63 Mbp detected for oil content in the $F_{2:3}$ population in the present study. Recently, cqSeed protein-003 located on chromosome 20 was narrowed down through fine-mapping to a 77.8 kb region between genetic marker BARCSOYSSR_20_0670 and BARCSOYSSR_20_0674 (31.74 to 31.82 Mbp), and the *Glyma.20G85100* gene encoding the CCT domain was identified as a candidate gene involved in protein content [20].

In our results, protein-related QTLs were mapped to Gm20_29512680 at 30.61 Mbp and Gm20_27578013 at 28.69 Mbp on chromosome 20 in the $F_{2:3}$ and $BC_1F_{2:3}$ populations, respectively. These results are consistent with the 24.55–32.91 Mbp range reported by Bolon et al. (2010) [15] and the 28.7–31.1 Mbp range reported by Hwang et al. (2014) [13], but not with the 32.71–33.08 Mbp range identified by Vaughn et al. (2014) [14] and the 31.74–31.82 Mbp range found by Fliege et al. (2022) [20]. The reason for these inconsistencies could the low LD with the surrounding markers [20], and it is known that the wild soybean variety used as a parent in this study has a lower LD than cultivated soybean [28]. For more accurate confirmation of the location, crossovers around the QTL detected in $BC_1F_{2:3}$ were identified but could not be found, and it was concluded that a crossover occurred at markers Gm20_33049242 and Gm20_32603292 in the two $BC_1F_{3:4}$ lines, which was advanced one generation by selecting high-protein lines. Therefore, it was predicted that the protein-related gene is present in the region downstream of Gm20_32603292.

Based on the results of QTL mapping, InDels were then searched for in the candidate genes located at around 30 Mbp between the Williams 82 reference genome and GWS-1887. Interestingly, no mutation was detected in the Glyma.20G85100 gene of the CCT motif family protein, which was recently cloned as a major protein-related gene [20]. These results suggest that other major protein-related genes may be present in a similar genetic region. Wang et al. (2021) selected protein-related candidate gene *Glyma.20g088000* using DEG analysis via RNA-seq, and it was found that *Glyma.20g088000* had a significant difference in its sequence between the high-protein Nanxiadou 25 and low-protein Tongdou 11 varieties due to InDels [16]. Interestingly, *Glyma.20g088000* (S-adenosyl-L-methionine-dependent methyltransferase) was also selected as a candidate gene in the present study because small

and large InDels occurred within several regions of the gene. In addition, *Glyma.20g086900* (aldehyde dehydrogenase-related) and *Glyma.20g088400* (oxidoreductase, 2-oxoglutarate-Fe(II) oxygenase family protein) genes were selected by Lee et al. (2019) as a result of a GWAS for the soybean seed protein content from maturation groups I to IV [17]. These two genes were also identified as candidate genes in this study.

In the present study, it was confirmed that *Glyma.20g088000* and *Glyma.20g088400* had a large InDel in the 5′ first exon and a small InDel in the 3′ third exon, respectively (Figure 5). Nonsense mutations that create stop codons and frameshifts in which amino acids are rearranged can disrupt the function of a gene [29]. In one example, truncated polypeptides generated as a result of nonsense mutations resulted in the loss of anthocyanin pigments associated with the color of soybean flowers [30]. In particular, the stop codon in *Glyma.20g088000* is expected to greatly simplify the structure of the protein (Figure 6), thus it is likely to have a significant effect on the expression of its function. Although these candidate genes have potential functions in metabolism, the mechanisms of how they relate to seed composition require further study. In addition, the results collectively suggest that protein content may be regulated by the complex interaction of multiple genes located at around 30 Mbp on chromosome 20.

4. Materials and Methods

4.1. Plant Materials

In the present study, 180 F_3 and 90 BC_1F_4 populations derived from a cross between Daepung and GWS-1887 were analyzed. Daepung, which was used as the female, recurrent parent, is an elite Korean variety that is strongly resistant to disease and shattering and has high yields [31], while GWS-1887, which was used as the male parent, was selected from the core collection of wild soybean accessions from the Rural Development Administration (RDA) because it has a protein content of 50% or higher. In the summer of 2018, F_1 seeds were obtained from artificial crossbreeding in an experimental field at Chonnam University (Gwangju, 36°17′ N, 126°39′ E, Republic of Korea). The F_1 seeds were planted in a greenhouse during the 2018–2019 winter season to obtain F_2 seeds, with the generation then advanced from F_2 to F_3 in the summer of 2019. At the same time, F_1 seeds were backcrossed in the summer of 2019 to obtain BC_1F_1 seeds. The produced BC_1F_1 seeds were planted in a greenhouse during the 2019–2020 winter period to obtain BC_1F_2 seeds. Finally, in the summer of 2020, the BC_1F_2 generation was advanced and BC_1F_3 seeds were obtained.

4.2. Analysis of Protein and Oil Content

All harvested seed samples were dried at 40 °C for 7 d and then pulverized using a coffee grinder to produce 3 g each for subsequent analysis. The crude protein content was measured using the Kjeldahl method. Reagents required for digestion, distillation, and titration were prepared, including 0.1N hydrochloric acid, a decomposition accelerator (containing 10 g of potassium sulfate and 1 g of copper sulfate), 40% sodium hydroxide solution, and 1% boric acid solution with 100 mL and 70 mL of Bromocresol green and methyl red solutions, respectively. The sample solution was prepared by mixing 0.7–1.0 g of the ground seed and 7–8 g of the decomposition accelerator with 10 mL of sulfuric acid in a decomposition bottle. Digestion was carried out by heating the sample solution at a slow ramping rate until no visible bubbles remained and the solution became transparent. The solution was then analyzed using a Kjeltec 1030 Autoanalyzer (FOSS Tecator AB, Hogans, Sweden) following the manufacturer's instructions.

The crude oil content was measured using ether extraction. For this, an oil metering bottle was pre-dried at around 95–100 °C for 2 h followed by cooling in a desiccator for 30 min. Following this, 2–3 g of the sample wrapped in No. 2 filter paper was dried at the same temperature and for the same duration of time as the oil metering bottle. After drying, the sample was placed in a Soxtec 1043 instrument (FOSS Tecator AB, Hogans, Sweden), and subjected to a flow of ether at 80 °C for 8 h to extract the oil. The processed

ether was then collected in an oil metering bottle and subsequently dried (95–100 °C for 3 h) followed by cooling in a desiccator (40 min) and weighing. The oil content was determined by subtracting the weight of the empty oil metering bottle from the weight of the bottle containing the extract.

$$Crude\ oil(\%) = \frac{\text{Oil bottle weight after extraction} - \text{Oil bottle weight before extraction}}{\text{Sample weight}} \times 100$$

4.3. DNA Extraction and SNP Genotyping

Fresh leaf tissue was collected at the beginning of growth for DNA extraction and ground using liquid nitrogen in a mortar. Genomic DNA was isolated from 20 mg of lyophilized leaf tissue using a DNeasy Plant Mini Kit (QIAGEN, Valencia, CA, USA) according to the manufacturer's protocol. The quality and quantity of the extracted total DNA were verified using a Nano-MD UV-Vis spectrophotometer (Scinco, Seoul, Republic of Korea). The extracted DNA was stored in a freezer at −80 °C until further use. A total of 270 samples, consisting of 180 F_2 and 90 BC_1F_2 plants and two replications of each parental plant (Daepung and GWS-1887) were genotyped using a SoySNP6K Illumina BeadChip (Illumina, San Diego, CA, USA) at TNT Research Co. (Anyang, Republic of Korea). The SNP alleles were called using Illumina's GenomeStudio (Illumina, Inc., San Diego, CA, USA).

4.4. Genetic Linkage Analysis

A genetic linkage map was constructed using the Kosambi mapping function in Joinmap v4.1 (Kyazma, Wageningen, The Netherlands). For genetic analysis, MQM mapping was employed with MapQTL 6.0 (Kyazma, Wageningen, The Netherlands). In the F2 population, permutations were conducted to determine the genome-wide significance threshold for the LOD scores, with the number of permutations set at 1000. In the BC1F2 population, an LOD score of $\geq$3.0 was set as the threshold for determining the presence of a QTL. LOD graphs and the location maps for the QTLs were created with MapChart2.2.

4.5. Re-Sequencing

Re-sequencing analysis was commissioned by Insilicogen (Yongin, Republic of Korea) and performed using an Illumine Novaseq 6000 platform. A library was constructed from DNA fragments with 151 bp paired ends read using a DNA Sample Prep Kit (Illumina) following the manufacturer's instructions. An analysis pipeline for detecting mutations in the sequencing data for the entire genome was employed with an NF-Core/SAREK workflow [32]. The snpEff tool was used for genetic variation annotation and effect prediction, while the snpEff database was referenced to Glycine max var. Williams 82 [11]. The whole genome sequencing data of GWS-1887 were deposited in NCBI under the BioProject PRJNA915129.

5. Conclusions

In this study, QTL mapping of the protein and oil content in soybean seeds was conducted using two progeny populations derived from high-protein wild soybean lines. QTL was detected in the region of *cqPRO*-003, which has been previously reported as a major QTL related to protein content, but as a result of resequencing, no difference was observed from the recently cloned candidate gene *cqPRO*-003. On the other hand, new candidate genes *Glyma.20g088000* and *Glyma.20g088400*, which contained InDels, were discovered. This suggests that the protein content may be regulated by the complex interaction of multiple genes and associations other than those that have previously been reported.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24044077/s1.

Author Contributions: Writing—original draft preparation, W.J.K.; methodology, B.H.K., C.Y.M., S.K. and S.S.; writing—review and editing, S.C.; resources, M.-S.C., S.-K.P. and J.-K.M.; supervision, B.-K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2018R1D1A1B07048126 and NRF-2015R1C1A1A02036757).

Data Availability Statement: The original contribution presented in the study are publicly available.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Natarajan, S.; Luthria, D.; Bae, H.; Lakshman, D.; Mitra, A. Transgenic Soybeans and Soybean Protein Analysis: An Overview. *J. Agric. Food Chem.* **2013**, *61*, 11736–11743. [CrossRef] [PubMed]
2. Kim, H.-K.; Kang, S.-T.; Choung, M.-G.; Jung, C.-S.; Oh, K.-W.; Baek, I.-Y.; Son, B.-G. Simple sequence repeat markers linked to quantitative trait loci controlling seed weight, protein and oil contents in soybean. *J. Life Sci.* **2006**, *16*, 949–954.
3. Kim, H.; Kang, S. Identification of Quantitative Trait Loci (QTLs) Associated with Oil and Protein Contents in Soybean (*Glycine max* L.). *J. Life Sci.* **2004**, *14*, 453–458. [CrossRef]
4. Wilcox, J.R.; Cavins, J.F. Backcrossing High Seed Protein to a Soybean Cultivar. *Crop Sci.* **1995**, *35*, 1036–1041. [CrossRef]
5. Hu, G.; Chen, Q.; Liu, C.; Jiang, H.; Wang, J.; Qi, Z. Integration of major QTLs of important agronomic traits in soybean. In *Soybean–Molecular Aspects of Breeding*; Sudaric, A., Ed.; InTech: London, UK, 2011; pp. 81–118.
6. Wilcox, J. Breeding soybeans for improved oil quantity and quality. In *World Soybean Research Conference III: Proceedings*; CRC Press: Boca Raton, FL, USA, 2022.
7. Kuroda, Y.; Tomooka, N.; Kaga, A.; Wanigadeva, S.M.S.W.; Vaughan, D.A. Genetic diversity of wild soybean (*Glycine soja* Sieb. et Zucc.) and Japanese cultivated soybeans [*G. max* (L.) Merr.] based on microsatellite (SSR) analysis and the selection of a core collection. *Genet. Resour. Crop Evol.* **2009**, *56*, 1045–1055. [CrossRef]
8. Lee, J.D.; Yu, J.-K.; Hwang, Y.-H.; Blake, S.; So, Y.-S.; Lee, G.-J.; Nguyen, H.T.; Shannon, J.G. Genetic diversity of wild soybean (Glycine soja Sieb. and Zucc.) accessions from South Korea and other countries. *Crop Sci.* **2008**, *48*, 606–616. [CrossRef]
9. Kofsky, J.; Zhang, H.; Song, B.-H. The Untapped Genetic Reservoir: The Past, Current, and Future Applications of the Wild Soybean (Glycine soja). *Front. Plant Sci.* **2018**, *9*, 949. [CrossRef]
10. Chen, Y.; Nelson, R. Genetic variation and relationships among cultivated, wild, and semiwild soybean. *Crop Sci.* **2004**, *44*, 316–325. [CrossRef]
11. Schmutz, J.; Cannon, S.B.; Schlueter, J.; Ma, J.; Mitros, T.; Nelson, W.; Hyten, D.L.; Song, Q.; Thelen, J.J.; Cheng, J.; et al. Genome sequence of the palaeopolyploid soybean. *Nature* **2010**, *463*, 178–183. [CrossRef]
12. Ha, J.; Abernathy, B.; Nelson, W.; Grant, D.; Wu, X.; Nguyen, H.T.; Stacey, G.; Yu, Y.; Wing, R.A.; Shoemaker, R.C.; et al. Integration of the draft sequence and physical map as a framework for genomic research in soybean (*Glycine max* (L.) Merr.) and wild soybean (*Glycine soja* Sieb. and Zucc.). *G3* **2012**, *2*, 321–329. [CrossRef]
13. Hwang, E.-Y.; Song, Q.; Jia, G.; E Specht, J.; Hyten, D.L.; Costa, J.; Cregan, P.B. A genome-wide association study of seed protein and oil content in soybean. *BMC Genom.* **2014**, *15*, 1. [CrossRef] [PubMed]
14. Vaughn, J.; Nelson, R.L.; Song, Q.; Cregan, P.B.; Li, Z. The Genetic Architecture of Seed Composition in Soybean Is Refined by Genome-Wide Association Scans Across Multiple Populations. *G3* **2014**, *4*, 2283–2294. [CrossRef] [PubMed]
15. Bolon, Y.-T.; Joseph, B.; Cannon, S.B.; A Graham, M.; Diers, B.W.; Farmer, A.D.; May, G.D.; Muehlbauer, G.J.; Specht, J.E.; Tu, Z.J.; et al. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biol.* **2010**, *10*, 41. [CrossRef] [PubMed]
16. Wang, J.; Mao, L.; Zeng, Z.; Yu, X.; Lian, J.; Feng, J.; Yang, W.; An, J.; Wu, H.; Zhang, M.; et al. Genetic mapping high protein content QTL from soybean 'Nanxiadou 25'and candidate gene analysis. *BMC Plant Biol.* **2021**, *21*, 1–13. [CrossRef]
17. Lee, S.; Van, K.; Sung, M.; Nelson, R.; LaMantia, J.; McHale, L.K.; Mian, M.A.R. Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theor. Appl. Genet.* **2019**, *132*, 1639–1659. [CrossRef] [PubMed]
18. 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]
19. Kim, M.; Schultz, S.; Nelson, R.L.; Diers, B.W. Identification and Fine Mapping of a Soybean Seed Protein QTL from PI 407788A on Chromosome 15. *Crop Sci.* **2016**, *56*, 219–225. [CrossRef]
20. Fliege, C.E.; 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.; et al. Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. *Plant J.* **2022**, *110*, 114–128. [CrossRef]
21. Fasoula, V.A.; Harris, D.K.; Boerma, H.R. Validation and Designation of Quantitative Trait Loci for Seed Protein, Seed Oil, and Seed Weight from Two Soybean Populations. *Crop Sci.* **2004**, *44*, 1218–1225. [CrossRef]

22. 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]
23. Leamy, L.J.; Zhang, H.; Li, C.; Chen, C.Y.; Song, B.-H. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genom.* **2017**, *18*, 18. [CrossRef] [PubMed]
24. Specht, J.; Chase, K.; Macrander, M.; Graef, G.; Chung, J.; Markwell, J.; Germann, M.; Orf, J.; Lark, K. Soybean Response to Water: A QTL Analysis of Drought Tolerance. *Crop Sci.* **2001**, *41*, 493–509. [CrossRef]
25. Boydak, E.; Alpaslan, M.; Hayta, M.; Gerçek, S.; Simsek, M. Seed Composition of Soybeans Grown in the Harran Region of Turkey As Affected by Row Spacing and Irrigation. *J. Agric. Food Chem.* **2002**, *50*, 4718–4720. [CrossRef]
26. Carrera, C.; Martínez, M.J.; Dardanelli, J.; Balzarini, M. Water Deficit Effect on the Relationship between Temperature during the Seed Fill Period and Soybean Seed Oil and Protein Concentrations. *Crop Sci.* **2009**, *49*, 990–998. [CrossRef]
27. 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]
28. Zhou, Z.; Jiang, Y.; Wang, Z.; Gou, Z.; Lyu, J.; Li, W.; Yu, Y.; Shu, L.; Zhao, Y.; Ma, Y.; et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **2015**, *33*, 408–414. [CrossRef] [PubMed]
29. Du, Y.; Luo, S.; Li, X.; Yang, J.; Cui, T.; Li, W.; Yu, L.; Feng, H.; Chen, Y.; Mu, J.; et al. Identification of Substitutions and Small Insertion-Deletions Induced by Carbon-Ion Beam Irradiation in Arabidopsis thaliana. *Front. Plant Sci.* **2017**, *8*, 1851. [CrossRef]
30. Takahashi, R.; Benitez, E.R.; Oyoo, M.E.; Khan, N.A.; Komatsu, S. Nonsense Mutation of an MYB Transcription Factor Is Associated with Purple-Blue Flower Color in Soybean. *J. Hered.* **2011**, *102*, 458–463. [CrossRef] [PubMed]
31. Park, K.; Kim, Y.H.; Lee, E.S.; Ha, K.S. A new soybean cultivar for fermented soyfood and tofu with high yield, "Daepung". *Korean J. Breed.* **2005**, *37*, 111–112.
32. Garcia , M.; Juhos, S.; Larsson, M.; Olason, P.I.; Martin, M.; Eisfeldt, J.; DiLorenzo, S.; Sandgren, J.; De Ståhl, T.D.; Ewels, P.; et al. Sarek: A portable workflow for whole-genome sequencin g analysis of germline and somatic variants. *F1000Research* **2020**, *9*, 63. [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.