



Article

Genome-Wide Association Study of Seed Quality and Yield Traits in a Soybean Collection from Southeast Kazakhstan

Botakoz Doszhanova ^{1,2} , Alibek Zatybekov ¹ , Svetlana Didorenko ³, Chao Fang ⁴, Saule Abugalieva ^{1,2}  and Yerlan Turuspekov ^{1,2,*} 

¹ Laboratory of Molecular Genetics, Institute of Plant Biology and Biotechnology, Almaty 050040, Kazakhstan; sybanbaeva_bota@mail.ru (B.D.); alexbek89@mail.ru (A.Z.); absaule@yahoo.com (S.A.)

² Faculty of Biology and Biotechnology, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan

³ Kazakh Research Institute of Agriculture and Plant Growing, Almaty 040909, Kazakhstan; svetl_did@mail.ru

⁴ School of Life Science, Guangzhou University, Guangzhou 510006, China; fangchao@gzhu.edu.cn

* Correspondence: yerlant@yahoo.com; Tel.: +7-7273948006

Abstract: Soybean (*Glycine max* (L.) Merr.) is a vital agricultural crop and a key source of protein and oil for food and feed production. The search for new genetic factors affecting the main agronomic traits of soybean is a significant step for efficient breeding strategies. This study aimed to identify marker–trait associations (MTAs) for seed protein and oil content and yield by conducting a genome-wide association study (GWAS). The collection of 252 soybean accessions of five different origins was analyzed over a period of five years. The GWAS was conducted using 44,385 SNP markers extracted from whole-genome resequencing data using Illumina HiSeq X Ten. The multiple-locus mixed linear model (MLMM) facilitated the identification of 38 stable MTAs: nine for protein content, nine for oil content, seven for the number of fertile nodes, six for the number of seeds per plant, four for thousand seeds weight, and three for yield per plant. Fifteen of these MTAs are presumed to be novel, with one linked to seed protein content, three linked to seed oil content, and the remaining MTAs linked to yield-related traits. These findings offer valuable insights for soybean breeding programs aimed at developing new, competitive cultivars with improved seed quality and yield characteristics.

Keywords: soybean; genome-wide association study; marker–trait association; protein content; oil content; yield components; genetic marker



Citation: Doszhanova, B.; Zatybekov, A.; Didorenko, S.; Fang, C.; Abugalieva, S.; Turuspekov, Y. Genome-Wide Association Study of Seed Quality and Yield Traits in a Soybean Collection from Southeast Kazakhstan. *Agronomy* **2024**, *14*, 2746. <https://doi.org/10.3390/agronomy14112746>

Academic Editors: Matthew Hegarty and Junhua Peng

Received: 4 October 2024

Revised: 4 November 2024

Accepted: 19 November 2024

Published: 20 November 2024



Copyright: © 2024 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.) Merr.) is a valuable plant oil and protein source for food and feed production. The seeds of soybean contain protein (~42%), carbohydrates (~33%), oil (~20%), and mineral nutrients (~5%) [1–5]. As of 2024, Brazil, the USA, and Argentina are the top soybean-producing countries in the world [6]. One of the regions where soybean cultivation is increasing at a rapid pace is Kazakhstan. In 2011, soybeans were cultivated on approximately 71,000 hectares of land in the country; by 2021, this area had tripled to over 200,000 hectares [7,8]. The climate and photoperiod in Kazakhstan limit the number of suitable areas for soybean cultivation; however, irrigation practices have enabled yields to reach up to 5.5 tons per hectare [9]. The primary regions for soybean cultivation in Kazakhstan are the Almaty region, which accounts for almost 90% of production, followed by East Kazakhstan and the Kostanay region [10]. The country is focusing on developing new soybean varieties with high productivity and protein and oil content to adapt to its diverse soil and climatic conditions [7,9,10].

Total soybean yield requires several important traits, including the number of fertile nodes, the number of seeds per plant, and the thousand seeds weight [11]. With regard to increasing yield, one of the main directions for soybean breeders is improving seed quality in new promising lines. Protein and oil content with yield components are complex

polygenic quantitative traits influenced by a combination of genetic and environmental factors, developmental processes, and trait interactions [12–14]. Only a few genes related to yield have been verified in soybean. For instance, *GmWRI1b* (the WRI from WRINKLED, the transcription factor regulating fatty acid biosynthesis) was found to increase fertile node number, seed number per plant, and yield per plant [15]. The gene *GmSWEET39* (Sugars Will Eventually be Exported Transporters) on chromosome 15 was shown to transport sucrose from the seed coat to the embryo and affect seed weight, size, and seed oil content in soybean [14,16]. The major QTL *cqSeed protein-003* on chromosome 20 has been reported to have the most significant additive effect of any protein QTL mapped in the crop, and *Glyma.20G85100* is the gene likely responsible for this important locus [12,17,18]. While progress has been made in identifying QTLs for soybean yield and quality traits, many underlying genes remain undiscovered and require further studies.

One of the challenges for high-quality soybean production is a negative correlation between protein content and oil/yield [5,12,17,19,20]. Therefore, different strategies can be applied to produce soybeans with higher-than-average protein genotypes and improved seed yield and oil content cultivars [21,22]. A promising approach to achieving this goal is using modern molecular genetics tools, such as marker-assisted selection. Genome-wide association study (GWAS) can be used in soybean improvement by identifying genetic markers associated with desirable traits to provide these markers for breeding programs [23–26]. Since the costs involved in whole-genome genotyping for accessions are gradually becoming increasingly accessible for the breeding community, GWASs are emerging as an effective technique for determining the genes underlying complex quantitative traits in soybean [27]. The first soybean GWAS was performed in around 2007–2008 and involved the use of SSR markers [28,29]. In 2012–2018, high-density single-nucleotide polymorphism (SNP) chips containing tens of thousands of markers were developed [30–32], including the SoySNP50K iSelect BeadChip from Illumina [33]. This chip contains over 50,000 SNPs and was one of the first high-density SNP chips developed for soybean, with it used for yield, agronomic traits, and seed composition [34,35]. The increasing accessibility of next-generation sequencing technologies (NGS) will likely lead to even denser SNP maps and a more complete picture of soybean genetic variation [36,37]. In this study, we aimed to identify marker–trait associations (MTAs) for yield components and protein and oil content in a diverse panel of soybean accessions through a genome-wide association study (GWAS) using whole-genome resequencing (WGRS) data. The panel was selected based on its genetic diversity, as the accessions originated from five different regions. The southeast region of Kazakhstan is the country’s main area of soybean production, and it is critical to understand the genetic basis of seed quality and yield to develop new soybean varieties for this region. The central hypothesis of this work is that a GWAS in this specific environment will facilitate the identification of additional MTAs to enhance local soybean breeding activities. In addition, accounting for different responses of the studied collections to given environmental factors on the identification of MTAs, the results may serve as a valuable reference for GWASs in other regions of the world.

2. Materials and Methods

2.1. Plant Material and Field Experiments

The soybean collection consisted of 252 breeding lines and cultivars from Kazakhstan ($n = 31$) and different countries in Eastern Europe ($n = 108$), Western Europe ($n = 23$), North America ($n = 40$), and East Asia ($n = 50$) (Table S1). The collection was grown in the experimental fields of the KRIAPG (Kazakh Research Institute of Agriculture and Plant Growing, Almaty region, Kazakhstan, 43°15' N, 76°54' W) from 2018 to 2022, with two replicates. Sowing and harvesting were performed from May to September based on Dospechov’s method [38]. The accessions were planted in four rows per plot: 25 cm plant spacing, 50 cm row spacing, and 1 m row length. The replicates were evaluated yearly in a randomized complete block design (RCBD) with randomly assigned soybean accessions. The accessions were grown under uncontrolled natural conditions without additional

treatment (fertilizers, fungicides, etc.). The field experiment design was standardized for all seasons. The mean daily temperature and precipitation at the KRIAPG were recorded during periods between key plant growth stages and are presented in Table S2.

2.2. Phenotyping of the Collection

Phenological observations were performed based on the method of Fehr and Caviness [39]. The planted soybean accessions were harvested at the R8 stage, the stage of full maturity, where 95% of the pods had reached their full mature color. Thereafter, the collection was screened based on yield component traits such as the number of fertile nodes (NFN, count), the number of seeds per plant (NSP, count), the thousand seeds weight (TSW, g), and the yield per plant (YP, g) based on the method of Korsakov and are presented in Table S1 [40]. The mean value of three random plants per plot was used for the phenotypic evaluation of each individual genotype. The soybean seeds were assessed for their seed protein content (SPC, %) and seed oil content (SOC, %) using the NIRS DS2500 Grain Analyzer (FOSS, Hillerød, Denmark), with the calibration supplied by the manufacturer. The analyzer operates by performing a sweep of frequencies, capturing data across a wide range of wavelengths from 700 to 2500 nm. We followed the manufacturer's recommendations throughout our analysis to ensure accuracy and reliability.

2.3. Statistical Analyses of Phenotypic Data

Descriptive statistics—minimum, maximum, mean, averages, standard error, and range values—for each year of the experiments were calculated. The single-factor analysis of variance (ANOVA) for every year with genotype (G), environment (R), and genotype \times environment (G \times E) interaction was calculated using SPSS v. 22 software (IBM Corp., Armonk, NY, USA), and broad-sense heritability (h^2) was estimated using the following formula:

$$h^2 = V_G / (V_G + V_E + V_{G \times E} + V_{\text{error}}), \quad (1)$$

where V_G is the genotypic variance, V_E is the environment variance, $V_{G \times E}$ is the variance due to the G \times E, and V_{error} is the error variance [41]. Genotypic variance was derived from the variation among the soybean accessions, whereas environmental variance was estimated from the variation observed across the five years of study. The correlation analysis of the studied traits was performed using SPSS v. 22.

2.4. Genotyping of the Collection

The WGRS data of 4,923,660 SNPs for 252 lines were obtained using the Illumina HiSeq X Ten system at the Department of the School of Life Sciences, Guangzhou University, China [42]. DNA samples were extracted and purified from the young leaves of individual cultivar single seeds using commercial kits (Qiagen, Redwood City, CA, USA). For each of the accessions in the panel, at least 5 μ g of DNA was used to construct a sequencing library with an Illumina TruSeq DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA), according to the manufacturer's instructions. Paired-end reads from the resequencing data were mapped to the reference genome of Chinese cultivar Zhonghuang 13 through the BWA software package (Cambridge, UK) [43]. The sequences close to indels were realigned with the IndelRealigner function using GATK v. 4.2 [44]. SNP and indel were identified using GATK v. 4.2 and SAMtools software (Boston, MA, USA) [42]. After excluding SNPs with more than 10% missing data and a minor allele frequency (MAF) of less than 1%, a total of 44,385 SNP markers were selected and used to determine the population structure and marker–trait association (Table S3).

2.5. The Assessment of Population Structure

The population structure of the studied soybean collection was analyzed using STRUCTURE v. 2.3.4 software [45]. STRUCTURE used a systematic Bayesian clustering approach applying Markov Chain Monte Carlo (MCMC) estimation [46]. We applied the admixture model with correlated allele frequencies, allowing for individuals to have ancestry from

multiple populations. The number of hypothetical groups ranging from $k = 1$ to 10 was assessed using 100,000 burn-in iterations followed by 100,000 recorded Markov-chain replications with the number of iterations equal to 3. The Delta K (ΔK), which determines the most appropriate number of clusters (K) in datasets, was identified based on the method of Evanno et al. [46] using the STRUCTURE HARVESTER v. 0.6.94 web-based program [47]. Files containing STRUCTURE results were used as input files for STRUCTURE HARVESTER, where we used the default settings to identify the most likely number of clusters. The obtained values were then transformed into a population structure (Q) matrix. The kinship matrix (K) that describes the most likely similarity of each allele between accessions was generated using TASSEL v. 5.0 [48]. Analysis of the linkage disequilibrium (LD) between the SNP markers based on their squared correlations (R^2) was also calculated in TASSEL and visualized in RStudio v. 2024.04.2 [49].

2.6. Genome-Wide Association Study and MTA–MTA Interactions

Associations between genotypic and single-trait phenotypic data were identified using the multi-locus mixed linear model (MLMM) with the GAPIT package (version 3) in RStudio (version 2024.04.2) [49,50]. The quantile–quantile (Q–Q) plots between the observed and expected \log_{10} p -values were compared to confirm the correction due to the K and Q matrices. The p -value $< 1 \times 10^{-4}$ was used as a significance threshold for the identified QTL, as the Bonferroni correction and false discovery rate (FDR) were too conservative and stringent for this analysis [51]. The SoyBase database [52] was used to search candidate genes for identified marker–trait associations with the reference genome of Chinese soybean Cv. Zhonghuang 13, genome assembly version 1 (glyma. Zh13. gnm1). The linkage disequilibrium decay distance was calculated by using the squared allele frequency correlation (r^2) in TASSEL and RStudio [49] and further applied as a boundary for identifying candidate genes near detected significant SNPs. To detect QTN-by-QTN interactions, we utilized the IIIVmrMLM method, which applies a multi-locus mixed linear model (MLMM) incorporating epistatic interactions between quantitative trait nucleotides (QTNs) to identify significant genetic associations influencing complex traits in soybean [53].

3. Results

3.1. Descriptive Statistics of Phenotypic Traits

The whole soybean collection was assessed using SPC and SOC and yield component traits, such as NFN, NSP, TSW, and YP, from 2018 to 2022 years. Over five years, the yield components showed significant diversity among the accessions. The mean value of the protein content ranged from 40.95% in 2019 to 43.29% in 2020. The oil content showed a small range of 20.31% to 21.01%. The maximum values of all yield components among the studied years were recorded in 2022, followed by 2018; minimum values were registered in 2019 for NFN, NSP, and TSW and in 2020 for YP. A summary of information on seed quality traits and yield components across five years of experiments is presented in Table 1.

A normal distribution was observed for all six traits, as illustrated in Figure S1. The check cultivar “Zhansaya” showed lower protein content values than the average datum for the entire soybean collection. For the oil content and YP, “Zhansaya” demonstrated values similar to the mean for the whole collection. For the remaining yield component traits (NFN, NSP, and TSW), the check cultivar showed lower values than the collection’s mean (Figure S1). The most promising genotypes identified in this study, which exhibited superior performance in both SPC ($>40\%$) and YP (>17 g), include the local breeding line 350/1 and the cultivars 1674 (China), Lybid (Ukraine), and Rainer 58 (Moldova) (Table S1).

ANOVA was used to assess the variance in the six studied traits based on genotype (G), environment/year, and genotype \times environment interaction (G \times E). The results of ANOVA, including p -values and heritability (h^2) of six traits, are presented in Table 2. The results showed that G, E, and G \times E interaction had a strong significant effect on all studied traits (p -value from $<2 \times 10^{-16}$ to 0.0106) except for the impact of G \times E on seed protein content (p -values 0.858). The heritability indices (h^2) were calculated for each trait (Table 2).

Table 1. Phenotypic variation in the seed protein and oil content and yield components of the soybean collection across the five years of study.

Traits	Value	2018	2019	2020	2021	2022	5-Year Mean
Seed protein content (SPC, %)	min	35.85	35.09	37.85	35.10	36.04	37.07
	max	48.59	48.22	48.60	49.30	47.70	48.36
	mean ± SE	41.67 ± 0.19	40.95 ± 0.19	43.29 ± 0.18	41.96 ± 0.19	42.33 ± 0.15	42.12 ± 0.16
Seed oil content (SOC, %)	min	16.45	15.87	16.60	16.20	16.89	17.05
	max	24.38	23.96	23.65	24.30	24.08	23.78
	mean ± SE	21.01 ± 0.10	20.87 ± 0.11	20.70 ± 0.11	20.97 ± 0.09	20.31 ± 0.09	20.70 ± 0.08
Number of fertile nodes (NFN, count)	min	6.90	4.90	4.60	4.70	4.00	8.32
	max	41.20	31.30	34.90	52.00	60.00	30.19
	mean ± SE	19.59 ± 0.41	14.45 ± 0.34	17.45 ± 0.35	15.87 ± 0.47	24.77 ± 0.66	18.42 ± 0.29
Number of seeds per plant (NSP, count)	min	15.00	8.40	6.70	8.00	5.80	17.04
	max	105.10	87.70	84.20	126.00	182.50	81.07
	mean ± SE	48.075 ± 1.12	36.51 ± 0.98	39.06 ± 0.86	40.27 ± 1.33	66.15 ± 1.84	46.03 ± 0.83
Thousand seeds weight (TSW, g)	min	124.00	18.90	11.90	118.00	114.00	106.12
	max	287.00	310.90	361.10	276.00	283.20	227.87
	mean ± SE	175.97 ± 1.51	125.97 ± 3.5	145.17 ± 4.99	173.52 ± 1.61	187.59 ± 1.83	161.56 ± 1.49
Yield per plant (YP, g)	min	3.00	1.50	0.40	0.30	1.12	3.82
	max	44.15	39.30	40.65	55.60	68.33	30.22
	mean ± SE	16.54 ± 0.46	10.50 ± 0.36	10.68 ± 0.34	9.11 ± 0.45	20.45 ± 0.73	13.45 ± 0.31

SPC—seed protein content; SOC—seed oil content; NFN—number of fertile nodes; NSP—number of seeds per plant; TSW—thousand seeds weight; YP—yield per plant; SE—standard error.

Table 2. ANOVA and heritability of seed quality and yield components in the studied soybean collection.

Trait	Factors	Df	SS	MS	p-Value	h ²
Seed protein content (SPC, %)	G	4	1091	272.87	<2 × 10 ⁻¹⁶ ***	0.64
	E	4	585	146.13	<2 × 10 ⁻¹⁶ ***	
	G × E	16	63	3.95	0.858	
	Res.	1051	6542	6.22		
Seed oil content (SOC, %)	G	4	182.0	45.49	<2 × 10 ⁻¹⁶ ***	0.65
	E	4	74.2	18.55	1.61 × 10 ⁻⁰⁷ ***	
	G × E	16	62.9	3.93	0.0106 *	
	Res.	1050	2057.9	1.96		
Number of fertile nodes (NFN, count)	G	4	2260	565	4.99 × 10 ⁻⁰⁹ ***	0.12
	E	4	16.120	4030	<2 × 10 ⁻¹⁶ ***	
	G × E	16	2105	132	0.000432 ***	
	Res.	1219	60.794	50		
Number of seeds per plant (NSP, count)	G	4	20.589	5147	7.91 × 10 ⁻¹¹ ***	0.12
	E	4	143.650	35.913	<2 × 10 ⁻¹⁶ ***	
	G × E	16	17.503	1094	0.000114 ***	
	Res.	1219	60.794	50		

Table 2. Cont.

Trait	Factors	Df	SS	MS	p-Value	h ²
Thousand seeds weight (TSW, g)	G	4	73.688	18.422	4.14×10^{-07} ***	0.10
	E	4	635.799	158.950	$<2 \times 10^{-16}$ ***	
	G × E	16	149.338	9334	7.42×10^{-09} ***	
	Res.	1221	2,499.565	2062		
Yield per plant (YP, g)	G	4	4168	1042	3.02×10^{-15} ***	0.15
	E	4	23.148	5787	$<2 \times 10^{-16}$ ***	
	G × E	16	2658	166	4.70×10^{-05} ***	
	Res.	1220	66.570	55		

SPC—seed protein content; SOC—seed oil content; NFN—number of fertile nodes; NSP—number of seeds per plant; TSW—thousand seeds weight; YP—yield per plant; Df—degree of freedom; SS—sum of squares; MS—mean of squares; h²—heritability index; ***— $p < 0.001$, *— $p < 0.05$.

The Pearson correlation analysis showed that the protein content and oil content were negatively and positively correlated with all studied traits ($p < 0.01$), respectively, whereas the remaining components were positively correlated (Table 3).

Table 3. Correlation coefficients among the studied traits (mean data of 2018–2022).

Traits	SPC	SOC	NFN	NSP	TSW
SOC	−0.632 **				
NFN	−0.513 **	0.289 **			
NSP	−0.573 **	0.314 **	0.913 **		
TSW	−0.413 **	0.384 **	0.407 **	0.476 **	
YP	−0.607 **	0.350 **	0.770 **	0.848 **	0.609 **

*** $p < 0.01$. SPC—seed protein content; SOC—seed oil content; NFN—number of fertile nodes; NSP—number of seeds per plant; TSW—thousand seeds weight; YP—yield per plant.

3.2. The Assessment of Population Structure

A total of 44,385 SNPs were used to analyze LD and population structure. The results of genome-wide LD decay analysis showed that at R² of 0.1, the distance was 321,509 bp (Figure S2). The population stratification was assessed using 44,385 SNP loci and the STRUCTURE package. The results produced by the STRUCTURE and STRUCTURE HARVESTER package applications suggested that K = 3 was the optimal value for the studied collection, as the correlation between K and Delta K illustrated the maximum Delta K at K = 3 (Figure 1A).

The distribution of soybean accessions among three generated subpopulations at K = 3 was as follows: Q1—80.5 % Eastern Europe, 8% Northern America, 5.5 % Western Europe, 3% East Asia, and 3% Kazakhstan; Q2—63% Eastern Europe, 19% Kazakhstan, 13% East Asia, 5% Northern America, and 0% Western Europe; Q3—29.2% Eastern Europe, 25.6% East Asia, 20.2% North America, 12.5% Kazakhstan, and 12.5% Western Europe (Figure 1B). Kazakhstan's soybean accessions were distributed among three subpopulations: 67.8% in Q3, 29% in Q2, and 3.2% in Q1.

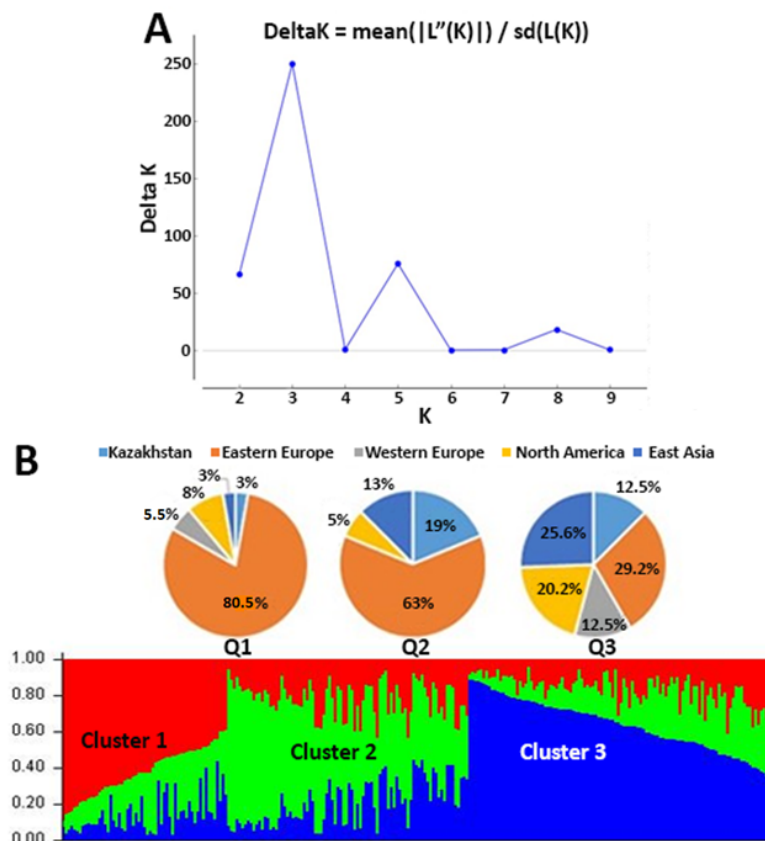


Figure 1. Population structure of the studied soybean collection. (A) Delta K (ΔK) plot and (B) Bayesian clustering of the 252 soybean accessions at $K = 3$.

3.3. Genome-Wide Association Study and MTA–MTA Interactions

The GWAS was conducted using 44,385 SNPs, field data for four yield-related traits, and protein and oil content. The results of the GWAS based on the MLM in GAPIT revealed 83 MTAs associated with the studied traits (Table 4). Among all 83 significant MTAs found in different environments, 38 were identified in at least two years of the experiment with significant p -values ($p < 1 \times 10^{-4}$) and considered stable MTAs. As a result, we identified nine associations for protein content, nine for oil content, seven for NFN, six for NSP, four for TSW, and three for YP (Table 4).

Table 4. The total number of marker–trait associations identified in this study.

Traits	Number of Significant QTLs *	Number of Stable MTAs	Number of Published MTAs **	Number of Novel MTA
SPC	15	9	8	1
SOC	12	9	6	3
NFN	16	7	4	3
NSP	15	6	3	3
TSW	14	4	1	3
YP	11	3	1	2
Total	83	38	23	15

* $p < 1 \times 10^{-4}$. ** The detailed information about MTAs with references is provided in Table S4. SPC—seed protein content; SOC—seed oil content; NFN—number of fertile nodes; NSP—number of seeds per plant; TSW—thousand seeds weight; YP—yield per plant.

The Manhattan and Q-Q plots of the GWAS for the mean values of five years of protein and oil content and for yield components are shown in Table S5. Detailed information about stable MTAs of the studied traits identified in at least two years of the experiment is presented

in Table 5. MTAs for protein content demonstrated a PVE range of 4.52–11.15% and the effect on the trait from -0.92 to $+0.92$. Among the MTAs for oil content, the largest PVE was 9.48% (SNP S16_6256537), and the largest negative effect was 1.29 (SNP S06_5537821).

Table 5. The list of identified significant MTAs associated with protein, oil, and yield components for 2018–2022 and the mean data.

Trait	SNP	Chr.	Position (bp)	<i>p</i> -Value	Positive Allele	Effect	PVE (%)	Year
SPC	S02_40916569	2	40.916.569	2.80×10^{-05}	C	-0.73	6.98	2018, mean
SPC	S05_3377432	5	3.377.432	5.06×10^{-05}	T	0.75	4.81	2020, mean
SPC	S06_5653082	6	5.700.406	1.92×10^{-05}	T	0.88	7.36	2018, 2021, 2022, mean
SPC	S06_48327505	6	48.327.505	2.66×10^{-05}	T	0.69	8.23	2020, 2022, mean
SPC	S10_45325143	10	45.325.143	3.67×10^{-05}	A	-0.92	7.24	2018, 2019, 2021, mean
SPC	S14_2073248	14	2.073.248	2.21×10^{-05}	A	-0.87	8.45	2019, 2020, mean
SPC	S14_45537479	14	45.537.479	5.53×10^{-05}	A	-0.85	4.52	2018, 2019, 2021, mean
SPC	S17_40642517	17	40.642.517	1.26×10^{-05}	A	-0.7	8.04	2022, mean
SPC	S19_48095581	19	48.015.856	1.04×10^{-06}	G	0.92	11.15	2018, 2021, mean
SOC	S03_3327118	3	3.349.008	4.03×10^{-05}	G	0.35	6.38	2019, 2020, 2022, mean
SOC	S06_5537821	6	5.653.082	2.04×10^{-05}	A	-1.29	5.56	2018, mean
SOC	S07_42119515	7	42.119.515	9.80×10^{-06}	G	-0.5	7.81	2018, 2022, mean
SOC	S09_41221274	9	41.221.274	7.71×10^{-05}	A	-0.49	4.97	2021, mean
SOC	S14_45542107	14	45.542.107	6.91×10^{-05}	T	0.59	6.94	2019, 2021, mean
SOC	S16_4427570	16	4.427.570	6.67×10^{-05}	G	0.55	6.76	2018, mean
SOC	S16_6256537	16	6.256.537	4.17×10^{-05}	T	0.59	9.48	2021, mean
SOC	S16_36471151	16	36.461.023	4.63×10^{-05}	A	-0.55	5.81	2018, 2021, mean
SOC	S17_5368514	17	5.368.514	8.13×10^{-05}	T	0.48	5.48	2021, mean
NFN	S07_17041516	7	16.978.432	6.77×10^{-06}	T	4.24	6.35	2022, mean
NFN	S11_8211474	11	8.211.474	6.01×10^{-05}	G	3.48	6.81	2022, mean
NFN	S13_28224204	13	28.224.204	2.28×10^{-06}	C	-3.68	7.50	2022, mean
NFN	S18_55651846	18	55.651.846	4.34×10^{-05}	G	1.44	7.01	2022, mean
NFN	S19_38590123	19	38.590.123	5.59×10^{-06}	T	4.37	8.12	2022, mean
NFN	S20_11932337	20	11.932.337	6.92×10^{-05}	C	-2.05	6.41	2019, 2022, mean
NFN	S20_34832847	20	34.832.847	1.37×10^{-05}	A	-1.36	9.02	2020, mean
NSP	S04_48109794	4	48.109.794	2.27×10^{-05}	G	13.54	6.32	2022, mean
NSP	S07_17041516	7	17.041.516	2.32×10^{-05}	T	10.72	5.08	2022, mean
NSP	S10_45323571	10	45.323.571	4.08×10^{-06}	T	5.59	8.27	2019, 2021
NSP	S16_8084401	16	8.084.401	5.45×10^{-05}	T	-9.19	4.97	2018, mean
NSP	S16_36572386	16	36.572.386	5.91×10^{-05}	C	6.94	6.78	2019, 2021, mean
NSP	S20_34832847	20	34.832.847	2.70×10^{-05}	A	-3.58	8.14	2020, mean
TSW	S06_3119005	6	3.119.005	8.94×10^{-05}	G	10.85	5.48	2020, mean
TSW	S10_5908937	10	5.908.937	8.83×10^{-05}	A	-9.66	4.61	2020, mean
TSW	S13_21819620	13	21.819.620	2.67×10^{-05}	T	16.22	6.54	2019, mean
TSW	S15_50483491	15	50.483.491	9.11×10^{-05}	C	-39.87	5.75	2020, mean
YP	S03_580652	3	580.652	4.16×10^{-05}	T	1.87	6.78	2019, 2021, mean
YP	S06_3459276	6	3.459.276	2.60×10^{-05}	C	-2.81	7.48	2018, 2020
YP	S08_14486855	8	14.486.855	4.08×10^{-09}	G	3.96	11.98	2019

SNP—single nucleotide polymorphism; SOC—seed oil content; SPC—seed protein content; NFN—number of fertile nodes; NSP—number of seeds per plant; TSW—thousand seeds weight; YP—yield per plant; PVE—phenotypic variants explained. Markers in bold are the novel MTAs for this study.

The MTAs for the NFN showed a PVE ranging from 6.35% to 9.02%; the most significant value was detected for SNP S20_34832847. For the NSP trait, the largest PVE was 8.27% (SNP S10_45323571), and the largest effect was +13.54 (SNP S04_48109794). MTAs associated with TSW exhibited PVE values from 4.61% to 6.54%, with trait effects ranging from −39.87 to +16.22. The most significant marker, S13_21819620 on chromosome 13, showed the highest PVE of 6.54% and the largest positive effect of +16.22. The marker S08_14486855 on chromosome 8 had the highest PVE at 11.98% and also the largest positive effect of 3.96. Haplotype analysis was conducted to identify combinations of genetic variants associated with seed quality and yield traits. The identified multiple haplotype blocks with row data for each trait are presented in Tables S6–S11. The weights of haplotypes with different numbers of effective alleles are shown in Figures S3–S8. The most effective haplotypes associated with increased protein and oil content for seed quality traits were the combination of alleles CCCTAAAGG (Table S6) and GATATGTAT (Table S7), respectively.

Notably, several SNPs in LD were associated with two traits simultaneously. For example, S16_36471151 and S16_36572386 were associated with both seed oil content and NSP. Similarly, S06_5653082, S06_5537821, S14_45537479, and S14_45542107 were in LD and associated with both protein and oil content. Additionally, S10_45325143 and S10_45323571 were associated with both protein content and NSP, and markers S07_17041516 and S20_34832847 were related to both NFN and NSP (Table 4).

Among the identified MTAs for seed quality and yield components, 23 MTAs matched with already reported QTLs from previous reports (Table 4). The interval of 321,509 bp (LD distance at $r^2 = 0.01$) was used as a window to map the candidate genes. Given the significant long-range LD, a wider distance of approximately 1.3 Mbp was allowed to present candidate genes, each annotated with its corresponding protein domain. Finally, we identified ten candidate genes, and their position overlapped with our identified MTAs (Table S4).

Specific genome-wide significant MTA–MTA interactions were detected for each studied trait except for YP (Table 6).

Table 6. Detected MTA–MTA interactions.

Trait	SNP	Chr.	Position (bp)	SNP	Chr.	Pos.	<i>p</i> -Value	lm_Coefficient	PVE (%)
SPC	S06_5653082	6	5.700.406	S06_48327505	6	48.327.505	5.45×10^{-04}	0.08725	0.54520
SPC	S14_2073248	14	2.073.248	S14_45537479	14	45.537.479	4.46×10^{-05}	−0.06032	0.06450
SPC	S19_48095581	19	48.015.856	S05_3377432	5	3.377.432	2.14×10^{-04}	0.06734	0.63980
SOC	S07_42119515	7	42.119.515	S09_41221274	9	41.221.274	4.87×10^{-04}	−0.05305	0.07854
SOC	S16_4427570	16	6.256.537	S16_6256537	16	36.461.023	5.40×10^{-05}	−0.05070	0.00965
NFN	S07_17041516	7	16.978.432	S11_8211474	11	8.211.474	2.10×10^{-05}	−0.05207	0.00874
NSP	S16_8084401	16	8.084.401	S10_45323571	10	45.323.571	4.50×10^{-04}	−0.04834	0.46540
NSP	S16_36572386	16	36.572.386	S20_34832847	20	34.832.847	8.78×10^{-05}	−0.04057	0.08748
TSW	S13_21819620	13	21.819.620	S15_50483491	15	50.483.491	2.46×10^{-04}	−0.03811	0.06885

SNP—single nucleotide polymorphism; SPC—seed protein content; SOC—seed oil content; NFN—number of fertile nodes; NSP—number of seeds per plant; TSW—thousand seeds weight; PVE—phenotypic variants explained.

In total, nine MTA–MTA interactions were detected at $p < 1 \times 10^{-03}$, including three interactions for SPC, two for SOC, two for NSP, one for NFN, and one for TSW. For YP, no significant interactions were found. The interactions captured a total of 1.249% of PVE in SPC, 0.088% of PVE in SOC, 0.009% of PVE in NFN, 0.553% of PVE in NSP, and 0.069% of PVE in TSW. Unlike the case for MTAs found in GWAS, no large- or moderate-effect MTA–MTA interactions were observed for studied traits.

4. Discussion

This research was conducted using 252 soybean accessions cultivated from 2018 to 2022 in the Almaty region, as this is the country's primary soybean production region [7,54].

Descriptive statistical analysis revealed significant variability for all studied traits, including SPC, SOC, NFN, NSP, TSW, and YP for the studied collection (Table 1 and Figure S1). Wide genetic diversity and a normal distribution within the analyzed population were sufficient for a successful GWAS and offered robust data for identifying genetic loci associated with the studied traits. The mean values of the phenotypic traits varied across the years, showing environmental factors' impact on the accessions' performance (Table 2). For instance, the mean SPC ranged from 40.95% in 2019 to 43.29% in 2020; in comparison, the SOC showed a relatively narrow range (Table 1). The ranges in yield components were also considerably variable, with the maximum values recorded in 2022 and the minimum recorded in 2019. High temperatures during the R5–R6 stage are known to decrease protein content but increase oil content [55]. In our study, the year 2020 was characterized by the coolest mean temperatures (Table S2), which may have resulted in slower lipid metabolism and greater protein accumulation (Table 1). In contrast, in 2018 and 2021, higher temperatures enhanced oil accumulation (Tables 1 and S2). A large number of studied accessions exceeded the performance of the check cultivar “Zhansaya” for all studied traits except SOC (Figure S1), indicating that the collection has great potential for breeding activities in the region. Thus, the variability of the soybean collection underscores the necessity of considering both genetic and environmental factors in breeding programs aimed at improving soybean traits [56]. Among the genotypes evaluated in this study, the most promising in terms of both SPC (>40%) and YP (>17 g) were the local breeding line 350/1 and the cultivars 1674 (China), Lybid (Ukraine), and Rainer 58 (Moldova). These genotypes demonstrated exceptional agronomic potential, making them valuable candidates for further breeding programs focused on enhancing soybean protein content and yield.

The ANOVA results indicated that G, E, and $G \times E$ significantly affected all the studied traits except for the $G \times E$ interaction on SPC and SOC (Table 2). The significant heritability indices (h^2) for SPC (0.64) and SOC (0.65) suggest that these traits are primarily influenced by genetic factors, making them promising targets for genetic improvement. However, the lower heritability indices for NFN (0.11), NSP (0.12), and TSW (0.10) indicate a more substantial environmental influence on these traits, highlighting the complexity of breeding for yield components [57]. The results of Pearson's correlation analysis revealed that SPC and SOC were negatively and positively correlated with all other traits (Table 3), respectively. This inverse relationship between SPC and SOC is critical for breeding programs and is recognized as evidence in genetics research, showing that both traits may compete for the same resources, are regulated by complex genetic networks, and have been historically selected for different uses, strengthening this inverse relationship [12,18,19,22,24,58].

High-yielding soybean cultivars are the main target of many world breeding programs to optimize productivity and meet global food demand [59]. In a recent GWAS, key genetic markers related to soybean yield components were identified through studies of diverse genotype panels [60,61]. In this study, the application of the MLM in a GWAS facilitated the identification of 83 significant MTAs of seed quality and yield-related traits, with 38 being stable across multiple years (Tables 4, 5 and S4). Among 38 stable MTAs, 20 yield-related associations were identified on 13 chromosomes across the soybean genome (Tables 4 and S4). The literature has previously reported nine regions [32,60,62–68]. For instance, in 2019, Karikari et al. [60] identified *qSW-19-3* associated with seed weight. In our study, in the close position, *q.NFN.ipbb.19* was associated with NFN, suggesting that this genome region may have a genetic factor with a pleiotropic effect. The gene *Rab5a2* encodes a small GTPase located close to MTA S13_28224204 and was discovered as a regulatory element for storage protein transport, influencing the seed protein content [63]. The MTAs *q.NFN.ipbb.20.2* and *q.NSP.ipbb.20*, associated with yield traits, are situated near the well-known protein gene *POWR1* on chromosome 20 [65,66]. The *POWR1* gene regulates lipid metabolism and nutrient transport in flowers and developing seed coats [65]. The gene *GmPDAT* [67], a regulator of chromosome condensation, affects oil content found in close position with TSW MTA, S13_21819620 (Table S4). The alignment of our findings with previously reported QTLs strengthens the robustness of our significance threshold.

Additionally, nine significant MTA–MTA interactions were identified for all traits except for YP (Table 6). However, their individual effects ranged from 0.009% to 0.640% of PVE, which is considerably smaller than the PVE values observed for MTAs in GWAS (ranging from 4.52% to 11.98%), indicating their limited contribution to the variance of studied traits.

Protein and oil content are crucial seed quality components and essential traits in soybean breeding. Although soybeans can achieve high yields, seed quality may often be compromised due to the negative genetic correlation between protein and oil content [34,69–71]. Recent studies have identified genetic loci that improve yield and seed quality, helping to develop more efficient soybean varieties [24,71,72]. This study identified nine MTAs associated with SPC spread across eight different chromosomes (Tables 5 and S4). The majority of these nine regions have already been reported in the scientific literature [2,70,73–75]. Two pairs of QTLs were identified for both SPC and SOC. These were *q.Prt.ipbb.6.1* and *q.Oil.ipbb.6* positioned on chromosome 6 at 5653082 and 5537821 bp, respectively, and *q.Prt.ipbb.14.2* and *q.Oil.ipbb.14* positioned on chromosome 14 at 45537479 and 45542107 bp, respectively (Tables 5 and S4). However, an opposite effect of minor alleles on SPC and SOC was observed, confirming a negative correlation between these two traits observed in the current study (Table 3) and in previous works [30,76]. Soybean gene *GA20OX* [77] positioned in the interval of 46961500–46963113 bp on chromosome 10 is close to MTAs *q.Prt.ipbb.10* and *q.NSP.ipbb.10* associated with SPC and NSP, respectively (Table S4). The gene is a key rate-limiting enzyme in gibberellins (GA) biosynthesis, producing bioactive GA [77]. In soybean, *GA20OX* is described as a key driver of seed traits and enhanced seed size and weight [78]. Jun Qin et al. (2022) [79] identified several QTLs for SPC on chromosomes 10 and 14 for SPC, findings that are consistent with our results (Table S4). In addition, Whiting et al. (2020) [73] reported several QTLs related to SPC, including *qPro_Gm02–3*, whose physical position is in the proximity of S02_40916569, identified in this study (Table S4). MTA with seed protein content *q.Prt.ipbb.6.2* was close to the candidate gene *GmZF351* (encoding a zinc finger protein) was identified [80]. This gene activates lipid biosynthesis pathways, contributing to increased oil accumulation in soybean seeds [80]. Additionally, *GmCYP78A72*, located in the vicinity of the MTA *q.Prt.ipbb.19*, is known to enhance seed size and weight [81]. The novel MTA *qPrt.ipbb.14.2* is located near the recognized membrane transport protein-like gene *SoyZH13_14G158400*, the component of the membrane regulating transport across the lipid bilayer [52] and could consequently be involved in controlling protein content through its impact on nutrient transport [82]. This consistency across different studies highlights the potential stability of these MTAs, making them strong candidates for marker-assisted selection.

This study identified nine MTAs for SOC on seven different chromosomes (Table 5). These MTAs showed substantial variability in their effects and the percentage of PVE, highlighting their importance in the genetic control of oil content in soybeans (Tables 4 and 5). The locations of MTAs on chromosomes 6, 9, 14, 16, and 17 overlapped with QTLs of seed oil content (Table S4) [35,66,75,83,84]. Two MTAs on chromosome 16 were identified in the same locations for QTLs for oil content identified by Jin and co-authors (2023) [66]. Moreover, S14_45542107 on chromosome 14 in this study was located near the QTL for oil content, as reported by Zhang and co-authors (2018) [35]. The gene *GmDof4* [84] was found to increase the content of lipids and seed weight in *GmDof4* transgenic Arabidopsis seeds. MTA *q.Oil.ipbb.17* was located close to this gene, suggesting that *q.Oil.ipbb.17* may regulate similar metabolic pathways influencing oil content and seed weight, as observed with *GmDof4*. The results of our literature survey suggest that three MTAs for SOC on chromosomes 3, 7, and 16 are putative novel factors for this trait. Among them, *q.Oil.ipbb.7* was found in coding DNA sequences, such as those involved in protein modification, gene *SoyZH13_07G202800* [52,85]. MTAs *q.Oil.ipbb.3* and *q.Oil.ipbb.16.1* were close to genes *SoyZH13_03G028300* and *SoyZH13_16G043200*, respectively [52]. These genes are involved in important biological functions such as leucine-rich repeat protein synthesis and cysteine protease activity [52]. The overlapping of these MTAs with genes with recognized functional roles suggests the possible involvement of these genes in the regulation of oil content in soybean.

Several identified MTAs in this report were associated with multiple traits, suggesting that they have pleiotropic effects. For example, S06_5653082 and S06_5537821 were linked to protein and oil content, indicating that these traits may share genetic regulation pathways and resources. Additionally, S16_36471151 and S16_36572386 were linked to both seed oil content and NSP and MTAs on chromosomes 7 and 20, which were pleiotropic for yield component traits with similar marker effects (Table 4). This pleiotropic effect is advantageous for breeding programs as they may simultaneously influence multiple agronomic traits, providing valuable targets for breeding programs.

To summarize the main findings, we identified 15 putatively novel MTAs for protein, oil, and yield components with no prior reports in the literature, emphasizing the discovery of potential new genetic factors contributing to soybean quality and yield traits.

5. Conclusions

A comprehensive assessment of a diverse collection of 252 soybean accessions, including seed quality and productivity traits, population structure, and MTAs, was performed. Significant variability in seed protein content (SPC), seed oil content (SOC), and yield components was observed across the studied accessions, offering valuable insights for breeding programs. The analysis confirmed the strong influence of genotype, environment, and genotype \times environment interaction on these traits, with high heritability indices for SPC and SOC. The GWAS identified 83 significant MTAs for key yield components and seed quality traits, 38 of which were stable across multiple years. Importantly, 15 novel MTAs were discovered, representing previously unreported genetic factors that could contribute to improving soybean yield and quality. Additionally, several MTAs demonstrated pleiotropic effects, influencing multiple traits such as SPC, SOC, and yield components. This highlights the potential for these markers to be used in marker-assisted selection (MAS) to simultaneously enhance several agronomic traits, making them highly valuable for future soybean breeding programs. In summary, this study advances our understanding of the genetic architecture of soybean quality and productivity traits and identifies novel genetic loci that could be targeted in breeding efforts to develop high-yielding, high-quality soybean varieties, particularly for Kazakhstan.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/agronomy14112746/s1>, Table S1: The list of the world soybean collection of 252 accessions with phenotypic data analyzed in this study; Table S2: The meteorological data in the Almaty region during the 5-year experiment period; Table S3: VCF file of 44385 SNP; Table S4: The list of GWAS-based identified marker–trait associations in soybean collection; Table S5: Manhattan and Q-Q plots for soybean seed quality and yield-related traits; Table S6: The row data for haplotype analysis of seed protein content; Table S7: The row data for haplotype analysis of seed oil content; Table S8: The row data for haplotype analysis of the number of fertile nodes; Table S9: The row data for haplotype analysis of the number of seeds per plant; Table S10: The row data for haplotype analysis of the thousand seeds weight; Table S11: The row data for haplotype analysis of the yield per plant; Figure S1: The normal distribution of studied traits; Figure S2: Decay of LD with physical distance in the soybean genome; Figure S3: Effect of the positive allele number on seed protein content in soybean; Figure S4: Effect of the positive allele number on seed oil content in soybean; Figure S5: Effect of the positive allele number on the number of fertile nodes; Figure S6: Effect of the positive allele number on the number of seeds per plant; Figure S7: Effect of the positive allele number on the thousand seeds weight; Figure S8: Effect of the positive allele number on the yield per plant.

Author Contributions: Conceptualization, A.Z. and Y.T.; methodology, A.Z. and Y.T.; validation, B.D., A.Z. and S.A.; formal analysis, B.D. and A.Z.; investigation, A.Z. and B.D.; resources, S.A., S.D. and C.F.; data curation, B.D., A.Z., S.A. and S.D.; writing—original draft preparation, B.D. and Y.T.; writing—review and editing, B.D., A.Z., C.F., S.A. and Y.T.; supervision, A.Z. and Y.T.; project administration, A.Z. and Y.T.; funding acquisition, A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education (former Ministry of Education and Science) of the Republic of Kazakhstan (Grant No. AP13068118 and Program No. BR24992903).

Data Availability Statement: The datasets generated and/or analyzed during the current study are available in the manuscript text and/or Supplementary Materials.

Acknowledgments: The authors acknowledge the technical assistance in the soybean seed quality analysis by the Biochemistry and Grain Quality Laboratory staff at the Kazakh Research Institute of Agriculture and Plant Growing, Almaty region, Kazakhstan.

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

References

- Bellaloui, N.; Reddy, K.N.; Bruns, H.A.; Gillen, A.M.; Mengistu, A.; Zobiolo, L.H.S.; Fisher, D.K.; Abbas, H.K.; Zablutowicz, R.M.; Kremer, R.J. Soybean Seed Composition and Quality: Interactions of Environment, Genotype, and Management Practices. In *Soybeans: Cultivation, Uses and Nutrition*; Nova Science Publishers: New York, NY, USA, 2011; pp. 1–42.
- 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](#)]
- Wijewardana, C.; Reddy, K.R.; Bellaloui, N. Soybean Seed Physiology, Quality, and Chemical Composition under Soil Moisture Stress. *Food Chem.* **2019**, *278*, 92–100. [[CrossRef](#)] [[PubMed](#)]
- Karr-Lilienthal, L.K.; Grieshop, C.M.; Spears, J.K.; Fahey, G.C. Amino Acid, Carbohydrate, and Fat Composition of Soybean Meals Prepared at 55 Commercial U.S. Soybean Processing Plants. *J. Agric. Food Chem.* **2005**, *53*, 2146–2150. [[CrossRef](#)]
- Medic, J.; Atkinson, C.; Hurburgh, C.R. Current Knowledge in Soybean Composition. *J. Am. Oil Chem. Soc.* **2014**, *91*, 363–384. [[CrossRef](#)]
- USDA Database. Available online: <https://www.usda.gov/> (accessed on 5 June 2024).
- Abugalieva, S.; Didorenko, S.; Anuarbek, S.; Volkova, L.; Gerasimova, Y.; Sidorik, I.; Turuspekov, Y. Assessment of Soybean Flowering and Seed Maturation Time in Different Latitude Regions of Kazakhstan. *PLoS ONE* **2016**, *11*, e0166894. [[CrossRef](#)]
- Zatybekov, A.; Yermagambetova, M.; Genievskaya, Y.; Didorenko, S.; Abugalieva, S. Genetic Diversity Analysis of Soybean Collection Using Simple Sequence Repeat Markers. *Plants* **2023**, *12*, 3445. [[CrossRef](#)]
- Yelnazarkyzy, R.; Kenenbayev, S.B.; Didorenko, S.V.; Borodychev, V.V. Soy cultivation technology with gravity drip irrigation in south and southeast Kazakhstan. *J. Ecol. Eng.* **2019**, *20*, 39–44. [[CrossRef](#)] [[PubMed](#)]
- Zatybekov, A.; Abugalieva, S.; Didorenko, S.; Rsaliyev, A.; Turuspekov, Y. GWAS of a Soybean Breeding Collection from South East and South Kazakhstan for Resistance to Fungal Diseases. *Vavilovskii Zhurnal Genet. Seleksii.* **2018**, *22*, 536–543. [[CrossRef](#)]
- Li, M.; Liu, Y.; Wang, C.; Yang, X.; Li, D.; Zhang, X.; Xu, C.; Zhang, Y.; Li, W.; Zhao, L. Identification of Traits Contributing to High and Stable Yields in Different Soybean Varieties Across Three Chinese Latitudes. *Front. Plant Sci.* **2020**, *10*, 1642. [[CrossRef](#)]
- 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](#)]
- Tian, X.; Zhang, K.; Liu, S.; Sun, X.; Li, X.; Song, J.; Qi, Z.; Wang, Y.; Fang, Y.; Wang, J.; et al. Quantitative Trait Locus Analysis of Protein and Oil Content in Response to Planting Density in Soybean (*Glycine max* [L.] Merri.) Seeds Based on SNP Linkage Mapping. *Front. Genet.* **2020**, *11*, 563. [[CrossRef](#)] [[PubMed](#)]
- Wang, S.; Liu, S.; Wang, J.; Yokosho, K.; Zhou, B.; Yu, Y.C.; Liu, Z.; Frommer, W.B.; Ma, J.F.; Chen, L.Q.; et al. Simultaneous Changes in Seed Size, Oil Content and Protein Content Driven by Selection of SWEET Homologues during Soybean Domestication. *Natl. Sci. Rev.* **2020**, *7*, 1776–1786. [[CrossRef](#)] [[PubMed](#)]
- Guo, W.; Chen, L.; Chen, H.; Yang, H.; You, Q.; Bao, A.; Chen, S.; Hao, Q.; Huang, Y.; Qiu, D.; et al. Overexpression of *GmWRI1b* in Soybean Stably Improves Plant Architecture and Associated Yield Parameters, and Increases Total Seed Oil Production under Field Conditions. *Plant Biotechnol. J.* **2020**, *18*, 1639–1641. [[CrossRef](#)]
- 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](#)]
- Diers, B.W.; Specht, J.E.; Graef, G.L.; Song, Q.; Rainey, K.M.; Ramasubramanian, V.; Liu, X.; Myers, C.L.; Stupar, R.M.; An, Y.Q.C.; et al. Genetic Architecture of Protein and Oil Content in Soybean Seed and Meal. *Plant Genome* **2023**, *16*, e20308. [[CrossRef](#)]
- 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](#)]
- Clemente, T.E.; Cahoon, E.B. Soybean Oil: Genetic Approaches for Modification of Functionality and Total Content. *Plant Physiol.* **2009**, *151*, 1030–1040. [[CrossRef](#)] [[PubMed](#)]
- 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](#)]

21. Eskandari, M.; Cober, E.R.; Rajcan, I. Genetic Control of Soybean Seed Oil: II. QTL and Genes That Increase Oil Concentration without Decreasing Protein or with Increased Seed Yield. *Theor. Appl. Genet.* **2013**, *126*, 1677–1687. [[CrossRef](#)]
22. Panthee, D.R.; Pantalone, V.R.; West, D.R.; Saxton, A.M.; Sams, C.E. Quantitative Trait Loci for Seed Protein and Oil Concentration, and Seed Size in Soybean. *Crop Sci.* **2005**, *45*, 2015–2022. [[CrossRef](#)]
23. Nakano, Y.; Kobayashi, Y. Genome-Wide Association Studies of Agronomic Traits Consisting of Field-and Molecular-Based Phenotypes. *Rev. Agric. Sci.* **2020**, *8*, 28–45. [[CrossRef](#)] [[PubMed](#)]
24. Li, X.; Shao, Z.; Tian, R.; Zhang, H. Mining QTLs and Candidate Genes for Seed Protein and Oil Contents across Multiple Environments and Backgrounds in Soybean. *Mol. Breed.* **2019**, *39*, 139. [[CrossRef](#)]
25. Liu, B.; Fujita, T.; Yan, Z.H.; Sakamoto, S.; Xu, D.; Abe, J. QTL Mapping of Domestication-Related Traits in Soybean (*Glycine max*). *Ann. Bot.* **2007**, *100*, 1027–1038. [[CrossRef](#)] [[PubMed](#)]
26. 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](#)]
27. Huang, X.; Han, B. Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annu. Rev. Plant Biol.* **2014**, *65*, 531–551. [[CrossRef](#)]
28. Jun, T.H.; Van, K.; Kim, M.Y.; Lee, S.H.; Walker, D.R. Association Analysis Using SSR Markers to Find QTL for Seed Protein Content in Soybean. *Euphytica* **2008**, *162*, 179–191. [[CrossRef](#)]
29. Funatsuki, H.; Hajika, M.; Hagihara, S.; Yamada, T.; Tanaka, Y.; Tsuji, H.; Ishimoto, M.; Fujino, K. Confirmation of the Location and the Effects of a Major QTL Controlling Pod Dehiscence, *QPDH1*, in Soybean. *Breed. Sci.* **2008**, *58*, 63–69. [[CrossRef](#)]
30. Hwang, E.Y.; Song, Q.; Jia, G.; Specht, J.E.; 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](#)]
31. Huang, X.; Zhao, Y.; Wei, X.; Li, C.; Wang, A.; Zhao, Q.; Li, W.; Guo, Y.; Deng, L.; Zhu, C.; et al. Genome-Wide Association Study of Flowering Time and Grain Yield Traits in a Worldwide Collection of Rice Germplasm. *Nat. Genet.* **2012**, *44*, 32–39. [[CrossRef](#)]
32. Zatybekov, A.; Abugalieva, S.; Didorenko, S.; Gerasimova, Y.; Sidorik, I.; Anuarbek, S.; Turuspekov, Y. GWAS of Agronomic Traits in Soybean Collection Included in Breeding Pool in Kazakhstan. *BMC Plant Biol.* **2017**, *17*, 63–70. [[CrossRef](#)]
33. Song, Q.; Hyten, D.L.; Jia, G.; Quigley, C.V.; Fickus, E.W.; Nelson, R.L.; Cregan, P.B. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. *PLoS ONE* **2013**, *8*, e54985. [[CrossRef](#)]
34. Diers, B.W.; Specht, J.; Rainey, K.M.; Cregan, P.; Song, Q.; Ramasubramanian, V.; Graef, G.; Nelson, R.; Schapaugh, W.; Wang, D.; et al. Genetic Architecture of Soybean Yield and Agronomic Traits. *G3 Genes Genomes Genet.* **2018**, *8*, 3367–3375. [[CrossRef](#)] [[PubMed](#)]
35. Zhang, J.; Wang, X.; Lu, Y.; Bhusal, S.J.; Song, Q.; Cregan, P.B.; Yen, Y.; Brown, M.; Jiang, G.L. Genome-Wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. *Mol. Plant.* **2018**, *11*, 460–472. [[CrossRef](#)]
36. Fallah, M.; Jean, M.; Boucher St-Amour, V.T.; O'Donoghue, L.; Belzile, F. The Construction of a High-Density Consensus Genetic Map for Soybean Based on SNP Markers Derived from Genotyping-by-Sequencing. *Genome* **2022**, *65*, 413–425. [[CrossRef](#)] [[PubMed](#)]
37. Zhang, H.; Jiang, H.; Hu, Z.; Song, Q.; An, Y.Q.C. Development of a Versatile Resource for Post-Genomic Research through Consolidating and Characterizing 1500 Diverse Wild and Cultivated Soybean Genomes. *BMC Genom.* **2022**, *23*, 250. [[CrossRef](#)]
38. Dospechov, B. *Methods of Field Experience*; Kolos: Moscow, Russia, 1979.
39. Fehr, W.R.; Caviness, C.E. Stages of soybean development. *Iowa State Univ. Coop. Ext. Serv. Spec. Rep.* **1977**, *80*, 1–3.
40. Korsakov, N.I.; Makashewa, R.H.; Adamova, O.P. *Methodical Instructions for Studying the Collection of Grain Legumes*; VIR: Leningrad, Russia, 1975; 59p.
41. Fehr, W.R. *Principles of Cultivar Development: Theory and Technique*; Macmillan Publishing Company: New York, NY, USA, 1991; Volume 1, pp. 247–258. [[CrossRef](#)]
42. Lu, S.; Dong, L.; Fang, C.; Liu, S.; Kong, L.; Cheng, Q.; Chen, L.; Su, T.; Nan, H.; Zhang, D.; et al. Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. *Nat. Genet.* **2020**, *52*, 428–436. [[CrossRef](#)]
43. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)] [[PubMed](#)]
44. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [[CrossRef](#)]
45. Porras-Hurtado, L.; Ruiz, Y.; Santos, C.; Phillips, C.; Carracedo, Á.; Lareu, M.V. An Overview of STRUCTURE: Applications, Parameter Settings, and Supporting Software. *Front. Genet.* **2013**, *4*, 98. [[CrossRef](#)]
46. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the Number of Clusters of Individuals Using the Software STRUCTURE: A Simulation Study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [[CrossRef](#)] [[PubMed](#)]
47. Earl, D.A.; vonHoldt, B.M. STRUCTURE HARVESTER: A Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conserv. Genet. Resour.* **2012**, *4*, 359–361. [[CrossRef](#)]
48. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics* **2007**, *23*, 2633–2635. [[CrossRef](#)] [[PubMed](#)]

49. Allaire, J. RStudio: Integrated Development Environment for R. *J. Wildl. Manag.* **2011**, *75*, 1753–1766.
50. Lipka, A.E.; Tian, F.; Wang, Q.; Peiffer, J.; Li, M.; Bradbury, P.J.; Gore, M.A.; Buckler, E.S.; Zhang, Z. GAPIT: Genome Association and Prediction Integrated Tool. *Bioinformatics* **2012**, *28*, 2397–2399. [[CrossRef](#)]
51. Kaler, A.S.; Purcell, L.C. Estimation of a Significance Threshold for Genome-Wide Association Studies. *BMC Genom.* **2019**, *20*, 618. [[CrossRef](#)]
52. Grant, D.; Nelson, R.T.; Cannon, S.B.; Shoemaker, R.C. SoyBase, the USDA-ARS Soybean Genetics and Genomics Database. *Nucleic Acids Res.* **2009**, *38*, 843–846. [[CrossRef](#)]
53. Li, M.; Zhang, Y.W.; Xiang, Y.; Liu, M.H.; Zhang, Z.C.; Zhou, Y.H.; Zuo, J.F.; Zhang, H.Q.; Chen, Y.; Zhang, Y.M. IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol. Plant.* **2022**, *15*, 1251–1253. [[CrossRef](#)]
54. Didorenko, S.V.; Araily Alenkhanovna, Z.A.; Sidorik, I.; Abuglieva, A.I.; Kudaibergenov, M.S.; Iskakov, A.R. Diversification of Crop Production by Means of Spreading Soybeans to the Northern Regions of the Republic of Kazakhstan. *Biosci. Biotechnol. Res. Asia* **2016**, *13*, 23–30. [[CrossRef](#)]
55. Nakagawa, A.C.; Ario, N.; Tomita, Y.; Tanaka, S.; Murayama, N.; Mizuta, C.; Iwaya-Inoue, M.; Ishibashi, Y. High temperature during soybean seed development differentially alters lipid and protein metabolism. *Plant Prod. Sci.* **2020**, *23*, 504–512. [[CrossRef](#)]
56. De Toledo, J.F.F.; Arias, C.A.A.; De Oliveira, M.F.; Triller, C.; Miranda, Z.D.F.S. Genetical and Environmental Analyses of Yield in Six Biparental Soybean Crosses. *Pesqui. Agropecu. Bras.* **2000**, *35*, 1783–1796. [[CrossRef](#)]
57. Xavier, A.; Rainey, K.M. Quantitative Genomic Dissection of Soybean Yield Components. *G3 Genes Genomes Genet.* **2020**, *10*, 665–675. [[CrossRef](#)] [[PubMed](#)]
58. 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](#)]
59. Wilcox, J.R. Sixty Years of Improvement in Publicly Developed Elite Soybean Lines. *Crop Sci.* **2001**, *41*, 1711–1716. [[CrossRef](#)]
60. Karikari, B.; Chen Sh Xiao, Y.; Chang, F.; Zhou, Y.; Kong, J.; Bhat, J.A.; Zhao, T. Utilization of Interspecific High-Density Genetic Map of RIL Population for the QTL Detection and Candidate Gene Mining for 100-Seed Weight in Soybean Front. *Plant Sci.* **2019**, *10*, 1001. [[CrossRef](#)]
61. Bhat, J.A.; Adeboye, K.A.; Ganie Sh, A.; Barmukh, R.; Hu, D.; Varshney, R.K.; Yu, D. Genome-wide association study, haplotype analysis, and genomic prediction reveal the genetic basis of yield-related traits in soybean (*Glycine max* L.). *Front. Genet.* **2022**, *13*, 953833. [[CrossRef](#)]
62. Xavier, A.; Jarquin, D.; Howard, R.; Ramasubramanian, V.; Specht, J.E.; Brian, W.D.; Song, Q.; Cregan, P.B.; Nelson, R.; Mian, R.; et al. Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population. *G3 Genes Genomes Genet.* **2018**, *8*, 519–529. [[CrossRef](#)] [[PubMed](#)]
63. Wei, Z.; Pan, T.; Zhao, Y.; Su, B.; Ren, Y.; Qiu, L. The small GTPase *Rab5a* and its guanine nucleotide exchange factors are involved in post-Golgi trafficking of storage proteins in developing soybean cotyledon. *J. Exp. Bot.* **2020**, *71*, 808–822. [[CrossRef](#)]
64. Liang, Q.; Chen, L.; Yang, X.; Yang, H.; Liu, S.; Kou, K.; Fan, L.; Zhang, Z.; Duan, Z.; Yuan, Y.; et al. Natural variation of *Dt2* determines branching in soybean. *Nat. Commun.* **2022**, *13*, 6429. [[CrossRef](#)]
65. Goettel, W.; Zhang, H.; Li, Y.; Qiao, Z.; Jiang, H.; Hou, D.; Song, Q.; Pantalone, V.R.; Song, B.-H.; Yu, D.; et al. POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nat. Commun.* **2022**, *13*, 3051. [[CrossRef](#)]
66. Jin, H.; Yang, X.; Zhao, H.; Song, X.; Tsvetkov, Y.D.; Wu, Y.E.; Gao, Q.; Zhang, R.; Zhang, J. Genetic Analysis of Protein Content and Oil Content in Soybean by Genome-Wide Association Study. *Front. Plant Sci.* **2023**, *14*, 1182771. [[CrossRef](#)] [[PubMed](#)]
67. Liu, J.Y.; Zhang, Y.W.; Han, X.; Zuo, J.F.; Zhang, Z.; Shang, H.; Song, Q.; Zhang, Y.M. An evolutionary population structure model reveals pleiotropic effects of *GmPDAT* for traits related to seed size and oil content in soybean. *J. Exp. Bot.* **2020**, *31*, 6988–7002. [[CrossRef](#)]
68. Ravelombola, W.; Qin, J.; Shi, A.; Song, Q.; Yuan, J.; Wang, F.; Chen, P.; Yan, L.; Feng, Y.; Zhao, T.; et al. Genome-Wide Association Study and Genomic Selection for Yield and Related Traits in Soybean. *PLoS ONE* **2021**, *16*, e0255761. [[CrossRef](#)]
69. Clevinger, E.M.; Biyashev, R.; Haak, D.; Song, Q.; Pilot, G.; Saghai Maroof, M.A. Identification of Quantitative Trait Loci Controlling Soybean Seed Protein and Oil Content. *PLoS ONE* **2023**, *18*, e0286329. [[CrossRef](#)] [[PubMed](#)]
70. Zhang, K.; Liu, S.; Li, W.; Liu, S.; Li, X.; Fang, Y.; Zhang, J.; Wang, Y.; Xu, S.; Zhang, J.; et al. Identification of QTNs Controlling Seed Protein Content in Soybean Using Multi-Locus Genome-Wide Association Studies. *Front. Plant Sci.* **2018**, *871*, 1690. [[CrossRef](#)] [[PubMed](#)]
71. Priyanatha, C.; Torkamaneh, D.; Rajcan, I. Genome-Wide Association Study of Soybean Germplasm Derived from Canadian × Chinese Crosses to Mine for Novel Alleles to Improve Seed Yield and Seed Quality Traits. *Front. Plant Sci.* **2022**, *13*, 866300. [[CrossRef](#)]
72. Wang, X.; Jiang, G.L.; Green, M.; Scott, R.A.; Song, Q.; 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](#)]
73. Whiting, R.M.; Torabi, S.; Lukens, L.; Eskandari, M. Genomic Regions Associated with Important Seed Quality Traits in Food-Grade Soybeans. *BMC Plant Biol.* **2020**, *20*, 485. [[CrossRef](#)]
74. Bandillo, N.; Jarquin, D.; Song, Q.; Nelson, R.; Cregan, P.; Specht, J.; Lorenz, A. A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* **2015**, *8*, 1–13. [[CrossRef](#)]

75. Zhu, X.; Leiser, W.L.; Hahn, V.; Würschum, T. Identification of Seed Protein and Oil Related QTL in 944 RILs from a Diallel of Early-Maturing European Soybean. *Crop J.* **2021**, *9*, 238–247. [[CrossRef](#)]
76. Filho, M.; Destro, D.; Miranda, L.A.; Spinoso, W.A.; Carrão-Panizzi, M.A.; Montalván, R. Relationships among oil content, protein content and seed size in soybeans. *Braz. Arch. Biol. Technol.* **2001**, *44*, 23–32. [[CrossRef](#)]
77. Rieu, I.; Ruiz-Rivero, O.; Fernandez-Garcia, N.; Griffiths, J.; Powers, S.J.; Gong, F.; Linhartova, T.; Eriksson, S.; Nilsson, O.; Thomas, S.G.; et al. The gibberellin biosynthetic genes *AtGA20ox1* and *AtGA20ox2* act, partially redundantly, to promote growth and development throughout the Arabidopsis life cycle. *Plant J.* **2008**, *53*, 488–504. [[CrossRef](#)] [[PubMed](#)]
78. Duan, Z.; Li, Q.; Wang, H.; He, X.; Zhang, M. Genetic regulatory networks of soybean seed size, oil and protein contents. *Front. Plant Sci.* **2023**, *14*, 1160418. [[CrossRef](#)] [[PubMed](#)]
79. Qin, J.; Wang, F.; Zhao, Q.; Shi, A.; Zhao, T.; Song, Q.; Ravelombola, W.; An, H.; Yan, L.; Yang, C.; et al. Identification of Candidate Genes and Genomic Selection for Seed Protein in Soybean Breeding Pipeline. *Front. Plant Sci.* **2022**, *13*, 882732. [[CrossRef](#)]
80. Li, Q.-T.; Lu, X.; Song, Q.-X.; Chen, H.-W.; Wei, W.; Tao, J.-J.; Bian, X.-H.; Shen, M.; Ma, B.; Zhang, W.-K.; et al. Selection for a Zinc-Finger Protein Contributes to Seed Oil Increase during Soybean Domestication. *Plant Physiol.* **2017**, *173*, 2208–2224. [[CrossRef](#)]
81. Zhao, B.; Dai, A.; Wei, H.; Yang, S.; Wang, B.; Jiang, N.; Feng, X. Arabidopsis KLU homologue *GmCYP78A72* regulates seed size in soybean. *Plant Mol. Biol.* **2016**, *90*, 33–47. [[CrossRef](#)]
82. Ho, H.L. Functional Roles of Plant Protein Kinases in Signal Transduction Pathways during Abiotic and Biotic Stress. *J. Biodivers. Bioprospecting Dev.* **2015**, *2*, 147. [[CrossRef](#)]
83. Pathan, S.M.; Vuong, T.; Clark, K.; Lee, J.D.; Grover Shannon, J.; Roberts, C.A.; Ellersieck, M.R.; Burton, J.W.; Cregan, P.B.; Hyten, D.L.; et al. Genetic Mapping and Confirmation of Quantitative Trait Loci for Seed Protein and Oil Contents and Seed Weight in Soybean. *Crop Sci.* **2013**, *53*, 765–774. [[CrossRef](#)]
84. Wang, H.; Zhang, B.; Hao, Y.; Huang, J.; Tian, A.; Liao, Y.; Zhang, J.; Chen, S. The soybean Dof-type transcription factor genes, *GmDof4* and *GmDof11*, enhance lipid content in the seeds of transgenic Arabidopsis plants. *Plant J.* **2007**, *52*, 716–729. [[CrossRef](#)]
85. Hemsley, P.A. The Importance of Lipid Modified Proteins in Plants. *New Phytol.* **2015**, *205*, 476–489. [[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.