frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 16 June 2022
doi: 10.3389/fpls.2022.882732

Identification of Candidate Genes and Genomic Selection for Seed Protein in Soybean Breeding Pipeline

Jun Qin[1], Fengmin Wang[1], Qingsong Zhao[1], Ainong Shi[2], Tiantian Zhao[1], Qijian Song[3], Waltram Ravelombola[4], Hongzhou An[1], Long Yan[1], Chunyan Yang[1]* and Mengchen Zhang[1]**

[1] National Soybean Improvement Center Shijiazhuang Sub-Center, North China Key Laboratory of Biology and Genetic Improvement of Soybean, Ministry of Agriculture, Hebei Laboratory of Crop Genetics and Breeding, Cereal & Oil Crop Institute, Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang, China, [2] Department of Horticulture, University of Arkansas, Fayetteville, AR, United States, [3] Soybean Genomics and Improvement Lab, United States Department of Agriculture - Agricultural Research Service (USDA-ARS), Beltsville, MD, United States, [4] Department of Soil and Crop Sciences, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Deyue Yu,
Nanjing Agricultural University, China

Reviewed by:
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Hengyou Zhang,
Key Laboratory of Soybean Molecular Design and Breeding, Northeast Institute of Geography and Agroecology (CAS), China
Javaid Akhter Bhat,
Nanjing Agricultural University, China

***Correspondence:**
Ainong Shi
ashi@uark.edu
Chunyan Yang
chyyang66@163.com
Mengchen Zhang
zhangmengchendd@163.com

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: 24 February 2022
Accepted: 16 May 2022
Published: 16 June 2022

Citation:
Qin J, Wang F, Zhao Q, Shi A, Zhao T, Song Q, Ravelombola W, An H, Yan L, Yang C and Zhang M (2022) Identification of Candidate Genes and Genomic Selection for Seed Protein in Soybean Breeding Pipeline. Front. Plant Sci. 13:882732. doi: 10.3389/fpls.2022.882732

Soybean is a primary meal protein for human consumption, poultry, and livestock feed. In this study, quantitative trait locus (QTL) controlling protein content was explored *via* genome-wide association studies (GWAS) and linkage mapping approaches based on 284 soybean accessions and 180 recombinant inbred lines (RILs), respectively, which were evaluated for protein content for 4 years. A total of 22 single nucleotide polymorphisms (SNPs) associated with protein content were detected using mixed linear model (MLM) and general linear model (GLM) methods in Tassel and 5 QTLs using Bayesian interval mapping (IM), single-trait multiple interval mapping (SMIM), single-trait composite interval mapping maximum likelihood estimation (SMLE), and single marker regression (SMR) models in Q-Gene and IciMapping. Major QTLs were detected on chromosomes 6 and 20 in both populations. The new QTL genomic region on chromosome 6 (Chr6_18844283–19315351) included 7 candidate genes and the Hap.X^{AA} at the Chr6_19172961 position was associated with high protein content. Genomic selection (GS) of protein content was performed using Bayesian Lasso (BL) and ridge regression best linear unbiased prediction (rrBULP) based on all the SNPs and the SNPs significantly associated with protein content resulted from GWAS. The results showed that BL and rrBLUP performed similarly; GS accuracy was dependent on the SNP set and training population size. GS efficiency was higher for the SNPs derived from GWAS than random SNPs and reached a plateau when the number of markers was >2,000. The SNP markers identified in this study and other information were essential in establishing an efficient marker-assisted selection (MAS) and GS pipelines for improving soybean protein content.

Keywords: *Glycine max*, genome-wide association study, genomic selection, genotyping by sequencing, protein content, single nucleotide polymorphism

INTRODUCTION

Soybean [*Glycine max* (L.) Merr.] provides about 60% of the vegetable-derived proteins worldwide and is a primary meal protein for human consumption, poultry, and livestock feed (Wolf, 1970; Patil et al., 2017). Improving protein content is one of the major breeding objectives in breeding programs (Li S. et al., 2019; Stewart-Brown et al., 2019). Traditional soybean breeding methods require phenotyping and multigeneration selection. Although molecular marker-assisted selection (MAS) by tagging the desired genes during breeding selection is an approach to make the selection more efficient (Collard et al., 2005), it is only relatively effective for traits with high heritability and controlled by major genes (Xu and Crouch, 2008; Xu et al., 2012; Patil et al., 2017). Genomic selection (GS) was developed for the selection of traits controlled by multiple genes, but it has not been practically applied due to the large variation of prediction accuracy in different populations and lacking efficient genotyping platforms (Zhang A. et al., 2017; Liu et al., 2018). With the rapid development of genomic tools and DNA sequencing technology, breeders and geneticists are able to explore molecular approaches to increase seed protein genetic gain (Song et al., 2004, 2013; Schmutz et al., 2010; Wang et al., 2020).

Linkage analysis (Hyten et al., 2004; Nichols et al., 2006; Pathan et al., 2013; Teng et al., 2017; Whiting et al., 2020) and genome-wide association study (GWAS) are powerful tools to identify markers associated with seed protein content in soybean (Hwang et al., 2014; Leamy et al., 2017; Lee et al., 2019; Li S. et al., 2019); to date, a total of 262 loci have been reported through linkage analysis and 107 loci have been reported through GWAS (Patil et al., 2017; Gangurde et al., 2020) per SoyBase.[1] These loci were on all the chromosomes, especially chromosome (Chr.) 15 and Chr. 20 (see text footnote 1/). Among these, several quantitative trait loci (QTLs), such as *cqPro-20* on Chr. 20 and *cqPro-15* on Chr. 15, were confirmed based on a low error rate (lower than 0.01) and in different populations (Patil et al., 2017). More than 150 candidate genes have been suggested to control seed protein content in soybean (Zhang D. et al., 2017; Zhang J. et al., 2018; Zhang Y. et al., 2018; Li S. et al., 2019; Zhang et al., 2019; Wang et al., 2020). The most described genes affecting seed protein content were sugar efflux transporter SWEET39 (*Glyma15g05470*) and sugar efflux transporter SWEET24 (*Glyma08g19580*) (Wang et al., 2020).

The populations used for mapping protein content in the previous reports included pedigree-based F2 and F4:6 (Csanádi et al., 2001; Chapman et al., 2003), recombinant inbred lines (RILs) population (Qi et al., 2014; Hacisalihoglu et al., 2018), backcross population (Sebolt et al., 2000; Liang et al., 2010), multiline population (Brummer et al., 1997; Wang et al., 2014; Whiting et al., 2020), nested association mapping population (Gangurde et al., 2020), and natural population (Hwang et al., 2014; Bandillo et al., 2015; Li D. et al., 2019). Most studies used a single population, but some studies used two populations for QTL verification (Vaughn et al., 2014; Zhang D. et al., 2017; Zhang et al., 2019); a few studies analyzed QTL using both the linkage mapping and associate mapping methods (Zhang et al., 2019).

The annual wild soybean (*Glycine soja*) is an important resource to improve soybean (Lam et al., 2010; Yao et al., 2020). Therefore, the objectives of this study were to: (1) identify QTL conferring seed protein content in RILs derived from cultivated and wild soybeans; (2) identify single nucleotide polymorphism (SNP) markers associated with seed protein content in GWAS and candidate genes controlling the trait; and (3) assess the accuracy of GS base on different SNP sets, training population size, and statistical models.

MATERIALS AND METHODS

Plant Materials

Recombinant Inbred Line

A population of 180 F9-derived RILs was developed from a cross of Jidou12 (*Glycine max*) and Ye9 (*Glycine soja*). Jidou12 is a high-yield cultivar with a high protein content that is grown in Shandong Jiaodong Peninsula, Hebei Province, and south-central Shanxi. The seed protein content averaged 46.48% for Jidou12 and 48.78% for Ye9 on a dry weight basis.

Natural Population

A total of 284 soybean genotypes, including 250 accessions selected from germplasm collection by Dr. Lijuan Qiu's laboratory at the Chinese Academy of Agricultural Sciences and 34 cultivars from Hebei Province, were used for the protein association analysis (**Supplementary Table 1**). These genotypes were originally from 10 provinces in China (202, 67.5%) and 6 states in the United States (76, 30.1%), South Korea (3, 1.2%), and Japan (2, 0.8%).

Field Design

Field experiments were conducted at Shijiazhuang (114°83′E, 38°03′N) in Hebei Province in a randomized complete block design with three replications in 2008, 2010, 2019, and 2020. The plot size was 3 m × 6 m with six rows and 50 cm space between rows in all the trials. The planting density was 225,000 plants per ha. Each year, the plots were irrigated once at the seed-filling stage. Plants were harvested after 95% of the leaves were falling off. Ten plants were randomly chosen from the middle of the plot for indoor laboratory seed protein content analysis when 95% of plants in the plot were matured.

Statistical Analysis of Phenotypic Data

Seed protein content was quantified using Fourier transform-near IR spectroscopy (Bruker MPA, Karlsruhe, Germany) at the North China Key Laboratory of Biology and Genetic Improvement of Soybean, Ministry of Agriculture. Under the Quant 2 method of OPUS (https://www.bruker.com/en/products-and-solutions/infrared-and-raman/opus-spectroscopy-software/downloads.html) version 5.5 software (Bruker MPA, Karlsruhe, Germany), the samples' protein content data were calculated using the dry basis model (Yan et al., 2008). Each RIL and accession from each replication

[1] https://www.soybase.org/

of each environment were detected three times using about 100–150 dry seeds and the average was used for statistical analysis. Analysis of variance was performed using JMP® (https://www.jmp.com/en_us/home.html) Genomics 7 (Sall et al., 2017). The least-squares mean (LSM) of the protein content of each soybean genotype from JMP was used as the phenotypic data in the association mapping.

Genotyping by Sequencing and Single Nucleotide Polymorphism Discovery

Genomic DNA was extracted from leaves of soybean plants using the QIAGEN DNeasy Plant Mini Kit (250). DNA was digested using the restriction enzyme *Ape*KI following the genotyping by sequencing (GBS) protocol described by Elshire et al. (2011). The 90 bp pair-end sequencing of accessions was performed using an Illumina HiSeq 2000 machine at the Genetic Research Institute, Chinese Academy of Sciences. GBS data alignment, mapping, and SNP discovery were done using Short oligonucleotide analysis Package (SOAP) family software. An average of 3.26 M short reads for each accession was aligned to soybean whole-genome sequence (Wm82.a2.v1) using SOAPaligner/soap2. SOAPsnp version 1.05 was used for SNP calling (Li et al., 2009; Li, 2011). Approximately, a half-million SNPs were discovered among the 284 soybean germplasm accessions. The SNPs were filtered before genetic diversity and association analyses. Soybean accession with >5% missing SNP and the >2% heterozygous SNP genotypes was eliminated. After the SNP dataset was filtered to remove those SNPs with minor allele frequency (MAF) <2%, missing data >5%, and heterozygous genotype >25%, a total of 10,115 SNPs were used for genetic diversity and association analysis (**Supplementary Figure 1**).

Genetic Maps

The genetic maps were constructed with JoinMap 4.0 (Van Ooijen, 2006) when the threshold for the logarithm of odds (LOD) was 3.0 based on 180 F9 RILs. QTL analysis of protein content in the RIL population was performed using single-trait Bayesian interval mapping (BIM), single-trait multiple interval mapping (SMIM), single-trait composite interval mapping maximum likelihood estimator (SMLE), single marker regression (SMR) method of Q-gene software (Joehanes and Nelson, 2008) with inclusive composite interval mapping (ICIM, http://www.isbreeding.net) (Meng et al., 2015). Variance components, QTL heritability, and QTL effect for seed protein content were estimated by QTLNetwork version 2.1 based on the phenotypic data (Yang et al., 2008). Only the QTL, which was mapped in similar physical locations (<1,500 kb) on the same chromosomes based on the five methods, was defined as a reliable QTL. The selected SNP markers were further tested for their effect by variance analysis using JMP Pro 10 (Sall et al., 2017).

Population Genetic Diversity and Association Analysis

STRUCTURE is a program that uses Bayesian methods to analyze multilocus data in population genetics (Kaeuffer et al., 2007). This study used a hybrid model and an allelic variation occurrence non-correlative model to examine the population structure of soybean germplasm. The number of the subpopulation (K) was assumed to be between 1 and 12. Each K was run 10 times, the Markov Chain Monte Carlo (MCMC) length of the burn-in period was 20,000, and the number of the MCMC iterations after the burn-in was 50,000. Delta K was used to determine appropriate *K*-values (Earl and vonHoldt, 2012). Next, CLUMPP was used to integrate the STRUCTURE-generated results with the "repeat 1,000" parameter. In addition, two different association mapping models were used to analyze the association between the molecular markers and traits, the TASSEL general linear model (GLM-Q), and the mixed linear model (MLM) combining kinship with population structure (Q-matrix) (Yu et al., 2006; Bradbury et al., 2007).

Identification of Candidate Genes

Linkage disequilibrium (LD) analysis was performed in the regions with SNP significantly associated with protein content; SNPs with $r^2 > 0.5$ in a 1-Mb window were considered to be in one linkage disequilibrium (LD) block in the heterochromatic regions. Haplotype analysis was conducted on all the SNPs within the LD block containing significant loci. Two databases, namely, the SoyBase(see text footnote 1) and the *Arabidopsis* Information Resource,[2] were used for gene annotation and preliminary screening of candidate genes were determined by combined bioinformatics and statistics.

Genomic Selection

Ridge regression best linear unbiased prediction (rrBLUP) and Bayesian Lasso Regression (BLR) were used to predict genomic estimated breeding value (GEBV) in GS (Endelman, 2011; Legarra et al., 2011). The packages "rrBLUP" (Endelman, 2011) and "BGLR" (Pérez and de los Campos, 2014) containing the GS models rrBLUP and Bayesian Lasso (BL), respectively, were run in R software.

Prediction accuracy of seed protein was evaluated for different SNP sets, including 22 significant SNPs detected from GWAS, 22 random SNPs, 100 random SNPs, 250 random SNPs, 500 random SNPs, 1,000 random SNPs, 2,000 random SNPs, 5,000 random SNPs, and 10,115 SNPs. The effect of training population size on GS accuracy was investigated by conducting cross-validation at different levels with 100 replications for each cross-validation fold from two to ten.

RESULTS

Seed Protein Content Variations in Two Populations

The seed protein content of the 180 RILs showed a biased normal distribution, seed protein content ranged from 34.69 to 58.71, and the Coefficient of variation (CV) was 23.39% (**Supplementary Figure 2A**). The seed protein content of the 284 accessions showed a biased normal distribution, seed

[2]https://www.arabidopsis.org/

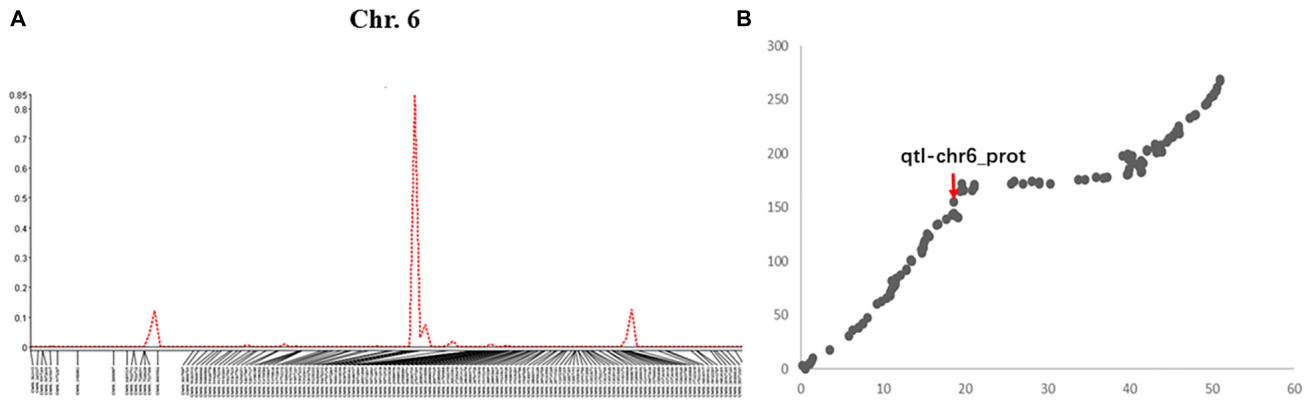

FIGURE 1 | (A) QTL mapping of seed protein content in soybean chromosome 6 based on single-trait multiple IM (SMIM) in Qgene, **(B)** The QTL, qtl-chr6_prot was mapped on the combined map between physical distance and genetic position of the chromosome 6, where the x-axis shows physical distance (Mbp) and the y-axis shows the genetic position (cM).

TABLE 1 | Single nucleotide polymorphism (SNP) markers/quantitative trait locus (QTL) detected in recombinant inbred line (RIL) and natural populations.

SNP Markers/QTL Detected in RIL and Natural populations	Population	Model	Confidence interval	Physical position bp	LOD	Posterior (POP)	PVE (%)
qtl-chr6_prot	RIL	Bayesian IM	142	18864382		0.847	
		Single-trait multiple IM (SMIM)	146–152	18580363–18597849	11.46		25.40
		Single-trait CIM MLE (SMLE)	142–152	18580363–18864382	13.1		
		Single marker regression (SMR)	141.9–144.5	18449510–19398117	13.7		29.60
		ICIM	144	18449510–18597849	14.11		22.30
Chr6_18658898	POP	MLM		18658898	19.95		
		GLM		18658898	25.76		
qtl-chr8_prot	RIL	Bayesian IM	56–58	9318625–9502316		0.392–0.49	
		Single-trait multiple IM(SMIM)	42–44	7270752–8285888	7.16		16.70
		Single-trait CIM MLE (SMLE)	42–44	7270752–8285888	7.05		
		Single marker regression	41.6–45.6	7270752–8285888	6.79		16.10
		ICIM	61	9701254–9877332	6.35		9.18
qtl-chr15_prot	RIL	Bayesian IM	12	1890050		0.179	
			32	4708800–4708818		0.759	
			42	5786875		0.119	
		Single-trait multiple IM (SMIM)	20–32	3303648–4708818	3.28		8.00
		Single-trait CIM MLE (SMLE)	18–20	3380704–3303648	4.43		
			30–50	4708800–6651199	4.93		
		Single marker regression	14.2–19.7	2095208–3303648	4.57		11.00
			27.7–31.8	4370908–4708800	3.91		9.50
			45.1–54.5	6037184–7193889	4.43		10.70
		ICIM	20	3303648–3488588	4.72		6.60
qtl-chr17_prot1	RIL	Bayesian IM	100	12398690–12801544		0.941	
		Single-trait multiple IM (SMIM)	100–124	12398690–13632893	4.11		9.80
		Single-trait CIM MLE (SMLE)	104–112	12801549–13813134	4.05		9.90
		Single marker regression	99.5–103.9	12398690–12801549	4.4		
qtl-chr20_prot	RIL	Bayesian IM	112	33202705		0.871	
		Single-trait multiple IM (SMIM)	94	33202705	6.31		14.90
		Single-trait CIM MLE (SMLE)	86–114	26572911–33224754	5.34		
		Single marker regression	93–115.3	26572981–33507017	5.12		12.30
		ICIM	97	26957096–27003724	7.16		10.22
Chr20_34423091	POP	MLM		34423091	7.21		
Chr20_34423091		GLM		34423091	6.55		

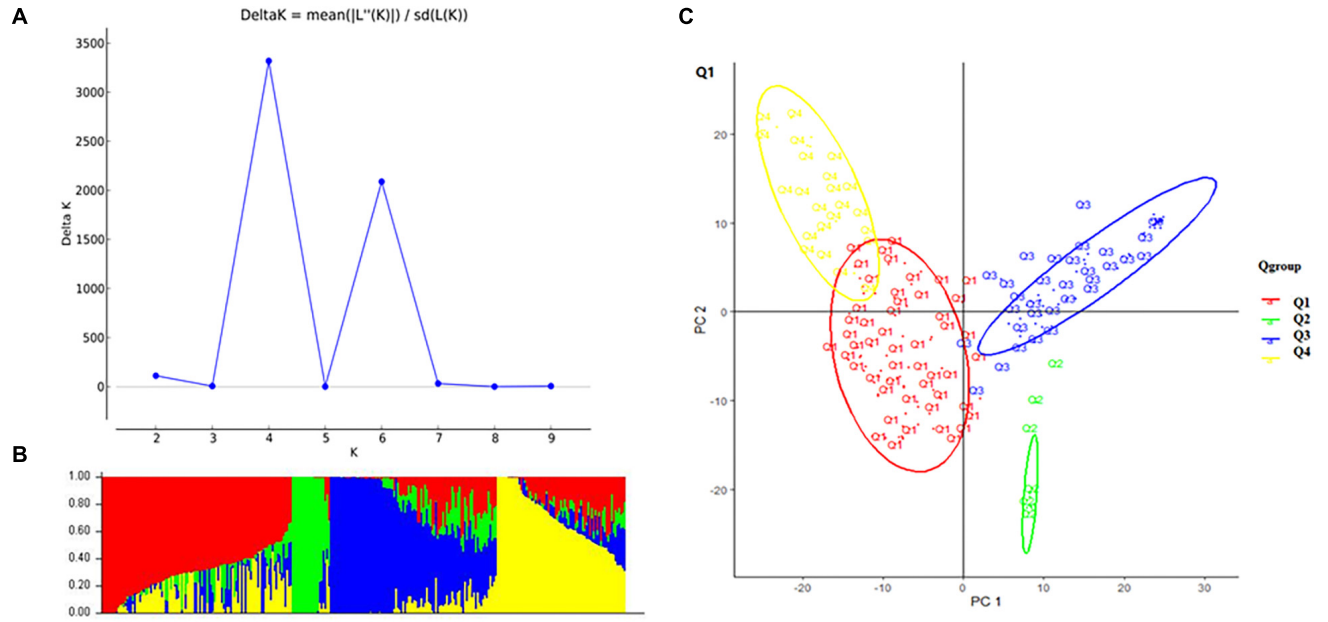

FIGURE 2 | Structure analysis: **(A)** delta *K*-values for different numbers of populations (K) from the STRUCTURE analysis, the *x*-axis shows different numbers of populations (K), the *y*-axis shows delta *K*-values for different numbers of subpopulations (K). **(B)** Classification of 284 accessions into four subpopulations using STRUCTURE version 2.3.4, where the *x*-axis shows accessions and the *y*-axis shows the probability (from 0 to 1) of each accession belonging to subpopulation (Q = K) membership. The membership of each accession belonging to subpopulations is indicated by different colors (Q1, red; Q2, green; Q3, blue; and Q4, yellow). **(C)** Principal component analysis (PCA) of the population structure. Distribution of the accessions in the association panel under PC1 and PC2.

protein ranged from 35.65 to 50.99, and the CV was 9.53% (**Supplementary Figure 2B**).

Genetic Map Construction and Quantitative Trait Locus Mapping in Recombinant Inbred Line Population

The RIL population was genotyped by sequencing. After filtering, a total of 2,498 polymorphic markers SNP were obtained and were mapped to 20 soybean chromosomes, thus the genetic maps were built for the RILs (**Supplementary Figure 3A**). According to their physical positions in the genome assembly, these markers were basically evenly distributed on 20 chromosomes. The 20 combined maps between physical distance and genetic position showed a good match (**Supplementary Figure 3B**). Chr. 14 had the least number of markers (68) and Chr. 18 had the largest number of markers (184). A genetic linkage map with a total length of 4,476.2 cm was constructed and the average distance between two adjacent markers was 1.8 cm (**Supplementary Figure 3**). The average distance between adjacent markers was the smallest on Chr. 20 (1.32 cm) but was the largest on Chr. 9 (2.26 cm).

A total of 5 QTLs on chromosomes 6, 8, 15, 17, and 20 were detected and the LOD value of the markers associated with the QTL ranged from 3.3 to 14.1; the QTL could explain 6.6%–29.6% of the genetic variation (**Figure 1A** and **Supplementary Figure 4**). Among these, one QTL with a positive allelic effect was from Jidou12 and 4 QTL with positive alleles were from Ye9 (**Table 1**). The QTL *qtl-chr6-prot* had the highest LOD and could explain 22.3–29.6% of genetic variation (**Table 1** and **Figure 1A**). The *qtl-chr6-prot* was in the heterochromatic region (**Figure 1B**).

Genome-Wide Association Study in Natural Population and Candidate Genes Selection

A total of 10,115 high-quality SNPs were used to perform population structure analysis of the 284 accessions using the STRUCTURE software (Kaeuffer et al., 2007). When $K = 4$, delta K was maximal with a relatively stable α value (**Figures 2A,B**). Cluster I was comprised of 102 accessions, including 77 cultivars, 21 landrace, and 5 exotic accessions; cluster II was comprised of 19 accessions, namely, 18 exotic accessions and 1 cultivar; cluster III was comprised of 93 accessions, namely, 57 exotic accessions, 34 cultivars, and 2 landraces; and cluster IV comprised of 70 accessions, namely, 51 landraces, 16 cultivars, and 3 exotic accessions. Principal component analysis (PCA) also showed the four groups (**Figure 2C**).

A significant association ($-\log P > 5.35$) with seed protein was observed for 22 SNPs from 22 haplotype blocks in 13 of the 20 chromosomes using GLM and MLM (**Table 2**). The LOD of the 22 markers ranged from 6.6 to 20.1 in GLM analysis and 6.3 to 26.3 in MLM analysis (**Table 2** and **Supplementary Figure 5**), indicating that these markers were strongly associated with seed protein. Eighteen of these markers were in euchromatic regions and four of these markers were in heterochromatic regions (**Table 2**).

TABLE 2 | Significant SNPs associated with protein content over 4 years, chromosome (Chr.) and physical position (bp) of the significant SNPs, logarithm of odds (LOD) [-log10 (*p*-value)] values of generalized linear model (GLM) and mixed liner model (MLM), and allele with positive effect at the SNP locus.

SNP Markers	Chr.	Position	Heterochromatic region	Euchromatic region	SNP Type	Allele with positive effect	LOD of GLM	LOD of MLM	SNP annotation
Chr03_34851073	3	34,851,073		E	A/C	C	12.79	13.69	Glyma.03G133300
Chr03_42692363	3	42,692,363		E	C/T	C	10.22	9.18	Glyma.03G224600
Chr05_40074496	5	40,074,496		E	A/T	T	20.03	25.86	Glyma.05G221300
Chr05_41114434	5	41,114,434		E	C/T	C	13.08	13.12	Upstream_gene_variant| MODIFIER| Glyma.05G234000
Chr06_14606307	6	14,606,307		E	A/G	G	9.30	8.41	Upstream_gene_variant| MODIFIER| Glyma.06G173600
Chr06_18658898	6	18,658,898	H		A/G	A	19.95	25.76	Glyma.06G202000
Chr08_10757609	8	10,757,609		E	C/T	C	7.23	6.65	Glyma.08G140700
Chr09_5898756	9	5,898,756		E	A/G	G	8.80	8.45	Glyma.09G062100
Chr09_45699847	9	45,699,847		E	A/G	G	8.55	7.64	Glyma.09G234500
Chr10_2992389	10	2,992,389		E	A/T	A	8.47	7.94	Glyma.10G034400
Chr10_44549078	10	44,549,078		E	A/G	G	9.22	8.80	Glyma.10G213000
Chr12_1536444	12	1,536,444		E	A/G	G	7.48	6.29	Glyma.12G021400
Chr14_2351357	14	2,351,357		E	C/T	C	10.58	9.26	Glyma.14G032300
Chr14_48312781	14	48,312,781		E	C/G	G	20.07	26.26	Upstream_gene_variant| MODIFIER| Glyma.14G218000
Chr15_13541492	15	13,541,492	H		C/G	C	8.48	7.71	Upstream_gene_variant| MODIFIER| Glyma.15G160000
Chr17_347445	17	347,445		E	A/T	A	9.38	9.40	Glyma.17G003000
Chr17_32480031	17	32,480,031	H		A/G	G	6.64	6.73	Intergenic_region| MODIFIER| Glyma.17G203300-Glyma.17G203400
Chr18_7837981	18	7,837,981		E	A/C	C	14.27	13.62	Glyma.18G081200
Chr18_18834295	18	18,834,295		E	A/C	C	8.24	7.25	Intergenic_region| MODIFIER| Glyma.18G133000-Glyma.18G133100
Chr18_50849168	18	50,849,168		E	A/T	A	11.29	11.66	Upstream_gene_variant| MODIFIER| Glyma.18G221300
Chr19_12210884	19	12,210,884	H		C/T	T	19.91	25.75	Intergenic_region| MODIFIER| Glyma.19G060900-Glyma.19G061000
Chr20_34423091	20	34,423,091		E	A/T	T	7.21	6.55	Glyma.20G100900

Two significant SNP loci on Chr. 6 and 20 were detected in linkage analysis and GWAS and the SNP loci detected on Chr. 6 by GWAS were in the QTL intervals obtained by linkage analysis. This SNP region on Chr. 6 had a high Phenotypic variation explained (PVE) (22.3–29.60%) and LOD (6.696–25.762). The region on Chr. 20 was associated with protein content with a PVE of 12.30% and LOD of 7.208 (**Tables 1**, **2**).

A 471-kb haplotype block from Chr6_18844283 to Chr6_19315351 included 7 SNP markers and 17 genes (**Figure 3A**). Pairwise LD analysis of the imputed SNP data showed that the candidate gene region was from Chr6_18842491 bp to Chr6_19015855 bp (**Figure 3B**). Seven candidate genes were in the regions, which included polynucleotidyl transferase (Glyma.06G202900 and Glyma.06G203100), polygalacturonase activity (Glyma.06G202600 and Glyma.06G203000), ATP synthase (Glyma.06G203200), and genes without annotation (Glyma.06G202700 and Glyma.06G202800) (**Figure 3B**).

There were 7 QTL haplotypes in the LD block from Chr6_18844283 to Chr6_19315351 in the natural population that showed differences in protein content (**Supplementary Table 2** and **Figure 3C**). The haplotypes Hap.B, Hap.C, and Hap.F had higher protein content than other haplotypes. Hap.B had the highest protein content, but no significant difference was observed among Hap.B, Hap.C, Hap.F, and Hap.G (**Figure 3C**). Further analysis showed that the SNP located at Chr6_19172961 may be more important; varieties carrying Hap.X^{AA} showed higher protein content than Hap.X^{GG} (**Figure 3D**).

Prediction Accuracy of Seed Protein Content

Prediction accuracy of different SNP densities for seed protein was conducted using 22 significant SNPs resulting from GWAS and 22 to 10,115 random SNPs, respectively. The prediction accuracy ranges from 0.44 to 0.77 using the rrBLUP model and from 0.44 to 0.78 using the BLR model (**Figure 4** and **Supplementary Table 3**). BLR and rrBLUP performed similarly for prediction accuracy; the average prediction accuracy was 0.63 and 0.53, respectively. The prediction accuracy of the 22 SNPs

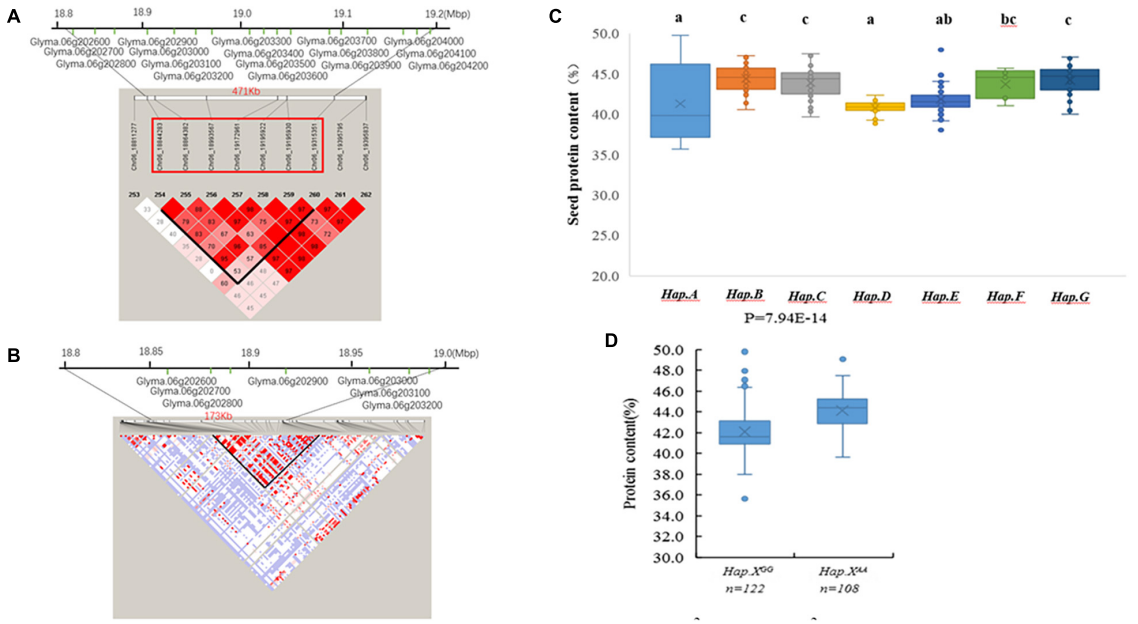

FIGURE 3 | (A) The extent of linkage disequilibrium (LD) in the regions based on pairwise r^2 values. The r^2 values are indicated using the color intensity index. Heatmap showing LD between each pair of markers that passed the Bonferroni threshold in genome-wide association study (GWAS). **(B)** Candidate genes for each single nucleotide polymorphism (SNP) locus. The bottom panel depicts the extent of linkage disequilibrium in the regions based on pairwise r^2 values. The r^2 values are indicated using the color intensity index shown. **(C)** Boxplot of seed protein based on different genotypes in soybean accessions. **(D)** Boxplot of seed protein based on Hap.X^{GG} and Hap.X^{AA} phenotypic differences between genotype combinations of the two SNPs.

obtained from GWAS was higher than that of random 22 SNPs and random 250 SNPs (**Figure 4** and **Supplementary Table 3**). Thus, regardless of the GS model, the accuracy of GS was higher when the significant SNPs from GWAS were used. Prediction accuracy for seed protein was increased with higher SNP density. However, there is a minimal difference in prediction accuracy after the SNP number reached 2,000 (**Figure 4** and **Supplementary Table 3**).

The effect of training population size on GS accuracy was also investigated by conducting cross-validation at different folds with 100 replications for each cross-validation (**Figure 5** and **Supplementary Tables 4**, **5**). On average, the prediction accuracy of the BLR model was 0.62 using GWAS-derived SNPs and 0.77 using the whole set of SNPs (**Figure 5** and **Supplementary Table 4**). The prediction accuracy of rrBLUP was less than BLR, with 0.5 using GWAS-derived SNPs and 0.77 using the whole set of SNPs (**Figure 5** and **Supplementary Table 5**). Considering average *r*-value and standardized deviation Sn, sevenfold resulted in a high *r*-value and low Sn in BLR models and sixfold resulted in a high *r*-value and low Sn in rrBLUP models.

DISCUSSION

Quantitative Trait Locus Mapping and Candidate Genes Identification for Soybean Seed Protein

Wild soybean with desired traits may improve the yield, quality, and other traits of cultivated soybeans. In this study, we performed QTL mapping for protein content in a RIL population derived from the cross of cultivated Jidou12 and wild soybean Ye9. Five major stable QTLs were detected on Chr. 6, 8, 15, 17, and 20 using Bayesian IM, SMIM, SMLE, and SMR models in Q-gene and IciMapping. Among these QTLs, we discovered that *qtl-chr6_prot* contributed an average of 25.77 of the phenotypic variance and the positive additive effects of allele were from the cultivated soybean Jidou12. The *qtl-chr6_prot* did not overlap with or was not adjacent to any of the previously reported QTLs for seed protein content. Other QTLs, *qtl-chr8_prot*, *qtl-chr15_prot*, *qtl-chr17_prot*, and *qtl-chr20_prot*, explained an average of 13.99, 9.1, 9.85, and 12.47 of the phenotypic variance, respectively; the positive additive effects of the allele of these QTL were from the wild soybean parent. The QTL *qtl-chr8_prot* (7.27–8.29 Mb) overlapped with the QTLs, as previously reported by Pathan et al. (2013). In addition, the QTL *qtl-chr15_prot* (3.30–4.71 Mb) overlapped with the *qPro15-1* (Zhang et al., 2019) and *qtl-chr17_prot* (12.80–13.81 Mb) with the protein 26-2 (Reinprecht et al., 2006). The position of QTL *qtl-chr20_prot* (26.57–33.51 Mb) was consistent with that of the confirmed QTL *cqPro-20* (Diers et al., 1992; Pandurangan et al., 2012; Vaughn et al., 2014; Sonah et al., 2015; Warrington et al., 2015; Zhang Y. et al., 2018; Fliege et al., 2022). Fliege et al. (2022) concluded that a transposon insertion within the CONSTANS, CO-like, and TOC1 (CCT) domain protein encoded by the *Glyma.20G85100* gene accounted for the high/low seed protein alleles of the cqSeed protein-003 QTL (31.74–31.84 Mb).

In the novel QTL region, the *qtl-chr6_prot*, seven candidate genes were identified. Of which, Glyma06G202900 and Glyma06G203100 were annotated as polynucleotidyl transferase,

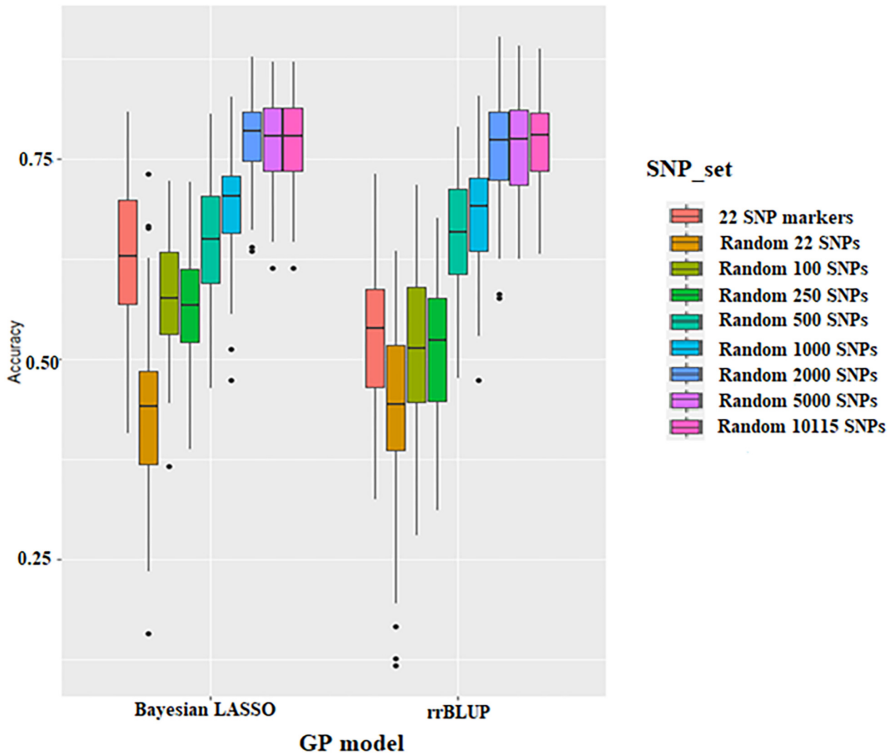

FIGURE 4 | Boxplots show the effect of different SNP density sets on genomic selection in the Bayesian Lasso Regression (BLR) model and ridge regression best linear unbiased prediction (rrBLUP) models.

ribonuclease H-like superfamily protein, which were homologous to the AT5G61090 gene in *Arabidopsis*. The protein encoded by the AT5G61090 had an RNA–DNA hybrid ribonuclease activity (Stoppel and Meurer, 2012). Glyma06G202600 was annotated as plasmodesmata callose-binding protein 3, homologous to AT1G18650 with callose-binding activity and the regulating intercellular trafficking in *Arabidopsis* (Simpson et al., 2009). Glyma06G203000 was annotated as a pectin lyase-like superfamily protein homologous to AT3G07820 with a polygalacturonase activity in *Arabidopsis* (Kim et al., 2006). Glyma06G203200 was annotated as a gamma subunit of Mt ATP synthase, homologous to AT2G33040, one of mitochondrial (mt) ATP synthesis subunits. Reduced expression of these subunits of the mt ATP synthase was proposed to disturb cellular redox states (Robison et al., 2009).

Genomic Selection in Soybean

Genomic selection overcomes the problems of traditional breeding methods and MAS selection and provides a new way for the selection of quantitative traits controlled by genes with minor effects. GS allows for the estimation of the effects of all the markers across the genome. These effects can be used to predict the performance of lines (Meuwissen et al., 2001). Since the target trait phenotype of an individual is predicted using the GS model, the materials could be screened and selected before planting, thus reducing costs and improving breeding efficiency (Heslot et al., 2012; Longin et al., 2015; Spindel et al., 2015). Matei et al. (2018) showed that the selection cycle for yield and seed weight can be significantly shortened using GS.

So far, the GS study has been mainly conducted on maize, wheat, and rice. The GS study in soybean remains limited. In 2013, Shu performed GS for 100-seed weight and reported a prediction accuracy of 0.904 (Shu et al., 2013). Subsequent GS showed accuracy for soybean cyst nematode (SCN) was 0.59–0.67 (Bao et al., 2014) and 0.64 for soybean yield (Jarquín et al., 2014).

The GS was performed on amino acid concentration (Qin et al., 2019), soybean chlorophyll content, soybean cyst nematode tolerance (Ravelombola et al., 2019), yield, and yield-related traits, such as maturity, plant height, and 100-seed weight (Ravelombola et al., 2021). These studies have shown the feasibility of GS for soybean yield and quality-related traits (Matei et al., 2018; Stewart-Brown et al., 2019).

However, few reports have focused on the GS of seed protein in soybean. Stewart-Brown et al. (2019) evaluated the potential of GS for soybean seed protein using 483 elite breeding lines from 26 biparentals and reported the predictive abilities of 0.81 in all the populations, 0.55 across populations, and 0.60 within each biparental population. Duhnen et al. (2017) compared genomic prediction accuracy of seed protein obtained using models calibrated across or within two subpopulations: early lines and late lines. The results showed that calibrations within subpopulations were more efficient. Five Bayesian models were also compared with Genomic best linear unbiased prediction (GBLUP) and did not show improved prediction accuracy. In this

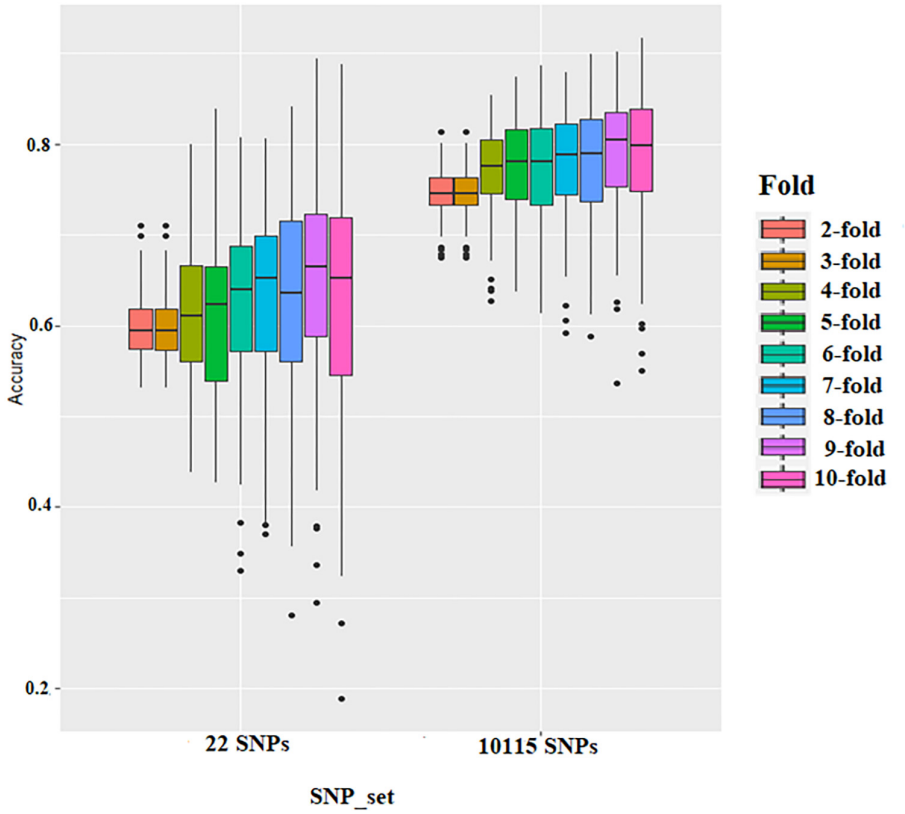

FIGURE 5 | Boxplots show the effect of training population size on genomic selection accuracy by conducting cross-validation at different folds with 100 replications for each cross-validation fold using rrBLUP.

study, we performed GS based on different SNP sets, different training population sizes, and statistical models. The results showed that the use of GWAS-derived SNPs for conducting GS significantly improved the accuracy of prediction, which was consistent with the results reported by Qin et al. (2019). The model selection criteria, SNP sets, and population training size were critical factors when conducting a GS, as reported in previous studies (Ravelombola et al., 2019, 2020, 2021). Those studies had demonstrated that 1,000–2,000 genome-wide markers across all the lines/accessions were needed to reach maximum efficiency of genomic prediction in the populations, increasing marker density that would not improve prediction efficiency (Poland et al., 2012; Bao et al., 2014; Zhang J. et al., 2016; Song et al., 2020). This study showed that there was a minimal difference in prediction accuracy after the SNP number reached 2,000 for seed protein content.

CONCLUSION

This study reported mapping and GS for seed protein content. Molecular markers associated with seed protein content were identified in RIL and natural populations and a novel QTL for seed protein content was detected and mapped on Chr. 6 in both populations. In addition, seven candidate genes that were related to seed protein content were identified. This is one of a few reports investigating seed protein content using RILs derived from cultivated and wild soybean crosses. Our results showed that GS accuracy was dependent on the SNP set and training population size; a set of GWAS-derived SNPs could increase GS accuracy. No significant GS accuracy difference was observed between rrBLUP and BL models. The results demonstrated the potential of using GS to improve soybean seed protein content.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

JQ and AS: data curation. JQ, CY, and LY: funding acquisition. FW, QZ, and TZ: investigation. JQ, AS, and WR: methodology. MZ, LY, and CY: project administration. AS: software. FW and TZ: validation. JQ: writing – original draft preparation. JQ, QS, and AS: writing - review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by: (1) the National Natural Science Foundation of China (32072092); (2) the Basic Research Funds of Hebei Academy of Agriculture and Forestry Sciences (2021060205); (3) the Special Innovation Program of Hebei Academy of Agriculture and Forestry Sciences (2022KJCXZX-LYS-6); (4) the S&T Program of Hebei, Soybean Modern Seed Industry Science and Technology Innovation Team (21326313D); (5) the Hebei Natural Science Foundation (2020301020); and (6) the China Agriculture Research System of The Ministry of Finance (MOF) and The Ministry of Agriculture and Rural Affairs (MARA) (CARS-04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.882732/full#supplementary-material

REFERENCES

Bandillo, N., Jarquin, D., Song, Q., Nelson, R. L., Cregan, P., Specht, J., et al. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome* 8, 1–13. doi: 10.3835/plantgenome2015.04.0024

Bao, Y., Vuong, T., Meinhardt, C., Tiffin, P., Denny, R., Chen, S., et al. (2014). Potential of association mapping and genomic selection to explore PI 88788 derived soybean cyst nematode resistance. *Plant Genome* 7, 2840–2854.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

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

Chapman, A., Pantalone, V., Ustun, A., Allen, F., Landau-Ellis, D., Trigiano, R., et al. (2003). Quantitative trait loci for agronomic and seed quality traits in an F 2 and F 4: 6 soybean population. *Euphytica* 129, 387–393.

Collard, B. C., Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* 142, 169–196.

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

Diers, B. W., Keim, P., Fehr, W., and Shoemaker, R. (1992). RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* 83, 608–612. doi: 10.1007/BF00226905

Duhnen, A., Gras, A., Teyssèdre, S., Romestant, M., Claustres, B., Daydé, J., et al. (2017). Genomic selection for yield and seed protein content in soybean: a study of breeding program data and assessment of prediction accuracy. *Crop Sci.* 57, 1325–1337.

Earl, D. A., and vonHoldt, B. M. (2012). Structure harvester: a website and program for visualizing structure output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255.

Fliege, C. E., Ward, R. A., Vogel, P., Nguyen, H., Quach, T., Guo, M., et al. (2022). Fine mapping and cloning of the major seed protein QTL on soybean chromosome 20. *Plant J.* 110, 114–128. doi: 10.1111/tpj.15658

Gangurde, S. S., Wang, H., Yaduru, S., Pandey, M. K., Fountain, J. C., Chu, Y., et al. (2020). Nested-association mapping (NAM)-based genetic dissection uncovers candidate genes for seed and pod weights in peanut (*Arachis hypogaea*). *Plant Biotechnol. J.* 18, 1457–1471. doi: 10.1111/pbi.13311

Hacisalihoglu, G., Burton, A. L., Gustin, J. L., Eker, S., Asikli, S., Heybet, E. H., et al. (2018). Quantitative trait loci associated with soybean seed weight and composition under different phosphorus levels. *J. Integr. Plant Biol.* 60, 232–241. doi: 10.1111/jipb.12612

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160.

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

Hyten, D. L., Pantalone, V. R., Sams, C., Saxton, A., Landau-Ellis, D., Stefaniak, T., et al. (2004). Seed quality QTL in a prominent soybean population. *Theor. Appl. Genet.* 109, 552–561. doi: 10.1007/s00122-004-1661-5

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., et al. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15:740. doi: 10.1186/1471-2164-15-740

Joehanes, R., and Nelson, J. C. (2008). QGene 4.0, an extensible Java QTL-analysis platform. *Bioinformatics* 24, 2788–2789. doi: 10.1093/bioinformatics/btn523

Kaeuffer, R., Réale, D., Coltman, D., and Pontier, D. (2007). Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity* 99, 374–380. doi: 10.1038/sj.hdy.6801010

Kim, J., Shiu, S. H., Thoma, S., Li, W. H., and Patterson, S. E. (2006). Patterns of expansion and expression divergence in the plant polygalacturonase gene family. *Genome Biol.* 7, 1–14. doi: 10.1186/gb-2006-7-9-r87

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F. L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059. doi: 10.1038/ng.715

Leamy, L. J., Zhang, H., Li, C., Chen, C. Y., and Song, B. H. (2017). A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genomics* 18:18. doi: 10.1186/s12864-016-3397-4

Lee, S., Van, K., Sung, M., Nelson, R., Lamantia, J., Mchale, L. K., et al. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theor. Appl. Genet.* 132, 1639–1659. doi: 10.1007/s00122-019-03304-5

Legarra, A., Robert-Granie, C., Croiseau, P., Guillaume, F., and Fritz, S. (2011). Improved Lasso for genomic selection. *Genet Res.* 93, 77–87. doi: 10.1017/S0016672310000534

Li, D., Zhao, X., Han, Y., Li, W., and Xie, F. (2019). Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. *Genomics* 111, 90–95. doi: 10.1016/j.ygeno.2018.01.004

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, S., Xu, H., Yang, J., and Zhao, T. (2019). Dissecting the genetic architecture of seed protein and oil content in soybean from the Yangtze and Huaihe River valleys using multi-locus genome-wide association studies. *Int. J. Mol. Sci.* 20:3041. doi: 10.3390/ijms20123041

Liang, H. Z., Yu, Y. L., Wang, S. F., Yun, L., Wang, T. F., Wei, Y. L., et al. (2010). QTL mapping of isoflavone, oil and protein contents in soybean (*Glycine max* L. Merr.). *Agr. Sci. China* 9, 1108–1116.

Liu, X., Wang, H., Wang, H., Guo, Z., Xu, X., Liu, J., et al. (2018). Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J.* 6, 341–352.

Longin, C. F. H., Mi, X., and Würschum, T. (2015). Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for

line and hybrid breeding. *Theor. Appl. Genet.* 128, 1297–1306. doi: 10.1007/s00122-015-2505-1

Matei, G., Woyann, L. G., Milioli, A. S., De Bem Oliveira, I., Zdziarski, A. D., Zanella, R., et al. (2018). Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. *Mol. Breed.* 38, 1–13.

Meng, L., Li, H. H., Zhang, L. Y., and Wang, J. K. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

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

Pandurangan, S., Pajak, A., Molnar, S. J., Cober, E. R., Dhaubhadel, S., Hernández-Sebastià, C., et al. (2012). Relationship between asparagine metabolism and protein concentration in soybean seed. *J. Exp. Bot.* 63, 3173–3184. doi: 10.1093/jxb/ers039

Pathan, S. M., Vuong, T., Clark, K., Lee, J. D., Shannon, J. G., Roberts, C. A., et al. (2013). Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. *Crop Sci.* 53, 765–774.

Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., et al. (2017). Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. *Theor. Appl. Genet.* 130, 1975–1991. doi: 10.1007/s00122-017-2955-8

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Poland, J. A., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113.

Qi, Z., Hou, M., Han, X., Liu, C., Jiang, H., Xin, D., et al. (2014). Identification of quantitative trait loci (QTL s) for seed protein concentration in soybean and analysis for additive effects and epistatic effects of QTL s under multiple environments. *Plant Breed.* 133, 499–507.

Qin, J., Shi, A., Song, Q., Li, S., Wang, F., Cao, Y., et al. (2019). Genome wide association study and genomic selection of amino acid concentrations in soybean seeds. *Front. Plant Sci.* 10:1445. doi: 10.3389/fpls.2019.01445

Ravelombola, W. S., Qin, J., Shi, A., Nice, L., Bao, Y., Lorenz, A., et al. (2019). Genome-wide association study and genomic selection for soybean chlorophyll content associated with soybean cyst nematode tolerance. *BMC Genomics* 20:904. doi: 10.1186/s12864-019-6275-z

Ravelombola, W. S., Qin, J., Shi, A., Nice, L., Bao, Y., Lorenz, A., et al. (2020). Genome-wide association study and genomic selection for tolerance of soybean biomass to soybean cyst nematode infestation. *PLoS One* 15:e0235089. doi: 10.1371/journal.pone.0235089

Ravelombola, W., Qin, J., Shi, A., Song, Q., Yuan, J., Wang, F., et al. (2021). Genome-wide association study and genomic selection for yield and related traits in soybean. *PLoS One* 16:e0255761. doi: 10.1371/journal.pone.0255761

Reinprecht, Y., Poysa, V. W., Yu, K., Rajcan, I., Ablett, G. R., and Pauls, K. P. (2006). Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. *Genome* 49, 1510–1527. doi: 10.1139/g06-112

Robison, M. M., Ling, X., Smid, M. P., Zarei, A., and Wolyn, D. J. (2009). Antisense expression of mitochondrial ATP synthase subunits OSCP (ATP5) and γ (ATP3) alters leaf morphology, metabolism and gene expression in Arabidopsis. *Plant Cell Physiol.* 50, 1840–1850. doi: 10.1093/pcp/pcp125

Sall, J., Stephens, M. L., Lehman, A., and Loring, S. (2017). *JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP*. Cary, NC: SAS Institute.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

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

Shu, Y., Yu, D., Wang, D., Bai, X., Zhu, Y., and Guo, C. (2013). Genomic selection of seed weight based on low-density SCAR markers in soybean. *Genet. Mol. Res.* 12, 2178–2188. doi: 10.4238/2013.July.3.2

Simpson, C., Thomas, C., Findlay, K., Bayer, E., and Maule, A. J. (2009). An Arabidopsis GPI-anchor plasmodesmal neck protein with callose binding activity and potential to regulate cell-to-cell trafficking. *Plant Cell* 21, 581–594. doi: 10.1105/tpc.108.060145

Sonah, H., O'donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* 13, 211–221. doi: 10.1111/pbi.12249

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One* 8:e54985. doi: 10.1371/journal.pone.0054985

Song, Q., Marek, L., Shoemaker, R., Lark, K., Concibido, V., Delannay, X., et al. (2004). A new integrated genetic linkage map of the soybean. *Theor. Appl. Genet.* 109, 122–128. doi: 10.1007/s00122-004-1602-3

Song, Q., Yan, L., Quigley, C., Fickus, E., Wei, H., Chen, L., et al. (2020). Soybean BARCSoySNP6K: an assay for soybean genetics and breeding research. *Plant J.* 104, 800–811. doi: 10.1111/tpj.14960

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., et al. (2015). Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11:e1004982. doi: 10.1371/journal.pgen.1004982

Stewart-Brown, B. B., Song, Q., Vaughn, J. N., and Li, Z. (2019). Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3* 9, 2253–2265. doi: 10.1534/g3.118.200917

Stoppel, R., and Meurer, J. (2012). The cutting crew–ribonucleases are key players in the control of plastid gene expression. *J Exp Bot.* 63, 1663–1673. doi: 10.1093/jxb/err401

Teng, W., Li, W., Zhang, Q., Wu, D., Zhao, X., Li, H., et al. (2017). Identification of quantitative trait loci underlying seed protein content of soybean including main, epistatic, and QTL× environment effects in different regions of Northeast China. *Genome* 60, 649–655. doi: 10.1139/gen-2016-0189

Van Ooijen, J. W. (2006). *JoinMap 4: Software for the Calculation of Genetic Linkage Maps in Experimental Populations.* Wageningen: Kyazma B.V.

Vaughn, J. N., Nelson, R. L., Song, Q., Cregan, P. B., and Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3* 4, 2283–2294. doi: 10.1534/g3.114.013433

Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y. C., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. *Natl. Sci. Rev.* 7, 1776–1786. doi: 10.1093/nsr/nwaa110

Wang, X., Jiang, G. L., Green, M., Scott, R. A., Song, Q., Hyten, D. L., et al. (2014). Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. *Mol. Genet. Genomics* 289, 935–949. doi: 10.1007/s00438-014-0865-x

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

Whiting, R. M., Torabi, S., Lukens, L., and Eskandari, M. (2020). Genomic regions associated with important seed quality traits in food-grade soybeans. *BMC Plant Biol.* 20:485. doi: 10.1186/s12870-020-02681-0

Wolf, W. J. (1970). Soybean proteins. Their functional, chemical, and physical properties. *J. Agric. Food Chem.* 18, 969–976.

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci.* 48, 391–407.

Xu, Y., Lu, Y., Xie, C., Gao, S., Wan, J., and Prasanna, B. M. (2012). Whole-genome strategies for marker-assisted plant breeding. *Mol. Breeding* 29, 833–854.

Yan, L., Jiang, C. Z., Yu, X. H., Yang, C. Y., and Zhang, M. C. (2008). Development and reliability of near infrared spectroscopy (NIS) models of protein and oil content in soybean. *Soybean Sci.* 27, 833–837.

Yang, J., Hu, C., Hu, H., Yu, R., Xia, Z., Ye, X., et al. (2008). QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics* 24, 721–723. doi: 10.1093/bioinformatics/btm494

Yao, Y., You, Q., Duan, G., Ren, J., Chu, S., Zhao, J., et al. (2020). Quantitative trait loci analysis of seed oil content and composition of wild and cultivated soybean. *BMC Plant Biol.* 20:51. doi: 10.1186/s12870-019-2199-7

Yu, J., Pressoir, G., Briggs, W. H., Vroh, Bi I, Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., et al. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Front. Plant Sci.* 8:1916. doi: 10.3389/fpls.2017.01916

Zhang, D., Lü, H., Chu, S., Zhang, H., Zhang, H., Yang, Y., et al. (2017). The genetic architecture of water-soluble protein content and its genetic relationship to total protein content in soybean. *Sci. Rep.* 7:5636. doi: 10.1038/s41598-017-04685-7

Zhang, J., Song, Q., Cregan, P. B., and Jiang, G. L. (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* 129, 117–130. doi: 10.1007/s00122-015-2614-x

Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B., et al. (2018). Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. *Mol. Plant* 11, 460–472. doi: 10.1016/j.molp.2017.12.016

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., et al. (2019). A combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. *Int. J. Mol. Sci.* 20, 1–19. doi: 10.3390/ijms20235915

Zhang, Y., He, J., Meng, S., Liu, M., Xing, G., Li, Y., et al. (2018). Identifying QTL–allele system of seed protein content in Chinese soybean landraces for population differentiation studies and optimal cross predictions. *Euphytica* 214, 1–17.

Conflict 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.