GENETIC IMPROVEMENT OF SEED COMPOSITION IN SOYBEAN

by

RENAN SILVA E SOUZA

(Under the Direction of Zenglu Li)

ABSTRACT

Soybean is an important protein source for human and animal nutrition. To improve the protein content, it is important to understand this trait genetics. A multiparent population with 1115 recombinant inbred lines (RILs) was evaluated for two years and genotyped with molecular markers aiming to introgress the 'Danbaekkong' chromosome 20 QTL and determine its effects. Based on published results, a marker targeting the gene *Glyma.20g085100* was developed to track the high protein allele. The QTL increased the protein by 3.3% on average, but yield was penalized. However, it was possible to identify lines with high protein and yield. The Danbaekkong allele was demonstrated to have originated from *Glycine soja* (PI 163453) and it is present in 79 *G. soja* accessions but absence in 35 *G. max* ancestors of North America cultivars. *G. soja* is a valuable source of alleles for protein improvement.

The second objective was to map protein and sulfur-containing amino acids QTLs. The RIL population Woodruff × PI 399000 was evaluated in six environments and genotyped with the SoySNP6K BeadChip. Three protein QTLs were identified on chromosomes (Chrs) 6, 15 and 17 and two QTLs on Chrs 6 and 10 for cysteine and methionine. The QTLs from PI 399000

increase protein without decreasing concentration of cysteine and methionine. Markers linked to the QTLs can be used to improve seed composition.

INDEX WORDS: Soybean, *Glycine max*, Chr 20 QTL, Seed composition, Protein, Amino acids, Yield, Quantitative trait loci (QTL) mapping, Multiparent population, *G. soja*, KASP markers.

GENETIC IMPROVEMENT OF SEED COMPOSITION IN SOYBEAN

by

RENAN SILVA E SOUZA

B.S. Federal University of Sao Joao del-Rei, 2014

M.S. University of Sao Paulo, 2018

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2023

GENETIC IMPROVEMENT OF SEED COMPOSITION IN SOYBEAN

by

RENAN SILVA E SOUZA

Major Professor: Zenglu Li
Committee: Esther Van Der Knaap
James W. Buck
Justin N. Vaughn
Paul L. Raymer

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2023

DEDICATION

For Sueme and Francisco – thank you for making every day special.

ACKNOWLEDGEMENTS

I am very thankful for all the amazing people that supported my work in completing this research. My advisor, Dr. Zenglu Li, has been a role model and a great mentor, providing guidance throughout the execution of projects and giving encouragement for my advancement. The committee members, Esther Van Der Knaap, James W. Buck, Paul L. Raymer, and Justin N. Vaughn have been essential in suggesting experiments and steering this research in the right direction. I am thankful to collaborators Drs. Blair Buckley and Benjamin Fallen for helping with the amino acid mapping project. I am thankful for the support provided by Carol Picard. Another important part of the completion of this research was the colleagues and friends at the Institute of Plant Breeding Genetics and Genomics. I am thankful for all the support provided by Niki Walden, Maddie Johnson, Deborah Franco, Stacey Gay. The soybean breeding and genetics team represented by Dale Wood, Brice Wilson, Earl Baxter, Gregory Gokalp, Brian Little, Tatyana Nienow, Nicole Bachleda and Breanna Sorg provided support for all my work and created a great environment. My fellow graduate students Ethan Menke, Alexandra Ostezan, Ivy Tran, Brooks Arnold, Samuel Mcdonald, Habib Widyawan, Nathaniel Burner, and Mark Miller gave me support in several aspects of my professional and personal life during my time in Athens. I am also thankful for the support from the United Soybean Board, the Coordination for the Improvement of Higher Education from Brazil (Capes), the John Ingle Innovation in Plant Breeding Award and the Glenn and Helen Burton Feeding the Hungry Scholarship Award.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Soybean Evolution, Diversity, and Importance

Soybean [*Glycine max* (L.) Merrill] belongs to the Fabaceae family, subfamily Papilionoideae, tribe Phaseoleae, genus Glycine. The genus *Glycine* is divided into two subgenera, *Glycine* (perennials) and *Soja* (annual) which includes *Glycine max* (Hymowitz & Newell, 1981). Similar to many cultivated species, soybean has a complex evolutionary history and research indicates that the ancestor of the genus Glycine (x = 10) underwent two genome duplication events approximately 59 and 13 million years ago (Schmutz et al., 2010). After that, a wild perennial ancestor (2n = 4x = 40) emerged and evolved to a wild annual (2n = 4x = 40) and finally to domesticated soybean (2n = 4x = 40). Despite the history of duplication events, all described species of the genus Glycine have normal diploid meiosis and soybean can be considered a paleopolyploid (Singh & Hymowitz, 1988). The first assembly of the reference genome (Williams 82 cultivar) was performed by Schmutz et al. (2010), which revealed the existence of approximately 46,000 genes positioned mostly in the chromosome ends and confirmed the polyploid origin of soybean cultivated today.

Modern soybean originated in China, and it is a product of the domestication that occurred approximately 3400 years ago. Cultivation expanded in Southeast Asia between the 15th and 16th centuries and in Europe in the early 18th century. In North America, the first records of cultivation date back to 1765 (Hymowitz & Harlan, 1983). The domestication process of soybean in Asia and introduction in North America have caused several genetic bottlenecks in

the species. Estimates indicate that approximately 80% of the rare alleles have been eliminated in modern soybean in comparison with wild ancestors (Hyten et al., 2006). This allele variation loss can be observed by the reduction in nucleotide diversity that occurred in each stage of the domestication process of soybean, in which the transition process of *Glycine soja* to *Glycine max* landraces caused a reduction from 2.17×10^{-3} SNPs/kb to 1.47×10^{-3} SNPs/kb. The transition of landraces to the North America ancestors further decreased the diversity to 1.14×10^{-3} SNPs/kb and the subsequent selection to generate elite cultivars reduced the diversity to 1.11×10^{3} SNPs/kb (Hyten et al., 2006).

The most severe cause of decrease in diversity was the initial process of domestication. The selection process applied on the North America ancestor lines to generate elite cultivars did not have a great impact on the diversity (Hyten et al., 2006; Valliyodan et al., 2016). Three possible explanations for this fact are given by Sedivy et al. (2017). Firstly, it is likely that the soybean breeders have unintentionally targeted traits associated with specific genomic regions containing the favorable alleles, which did not cause considerable changes in the genome landscape. Secondly, there has been an ongoing effort to broaden the genetic base of soybeans by introgressing alleles from *G. soja* and landraces which may have balanced the selection process. The third explanation is related to the soybean natural physiological variation in which different genotypes have different photoperiod requirements to reach maturity. As the crop expanded to different environments, there was a demand to develop cultivars with different maturity and these genotypes were distinct enough to curb further decrease in diversity.

After being used for centuries as an important ingredient in eastern Asian cuisine, soybean became a crop of great importance on a global scale in the 20th century. Soybean meal became the main source of protein for animal feed and soybean oil is now a key component in

human food, cosmetics, and biofuels. In 2021/2022 crop season, 28.5% of vegetable oil and 70% of protein meal used worldwide were derived from soybeans (USDA, 2023a). Over the past 50 years, world soybean production has increased approximately 10-fold, from 37.9 M tonnes in 1967 to 372 M tonnes in 2021 (FAO, 2023). The improvement in the standard of living in developing countries has resulted in an increase in the demand for meat and meat products, consequently increasing the need to produce soybean protein for animal feed. Another important factor is the increase in the demand for biofuels and the consequent growth in the production of vegetable oil (Qiu & Chang, 2010).

To meet the growing demand of farmers for higher yields and consumers for better quality of soybean products, plant breeding stands out as a key strategy because it can ensure improvements in yield and quality in a sustainable way. It has been demonstrated that the development of new soybean cultivars has ensured an annual increase of 13.7 kg ha^{-1} yr^{-1} in yield since 1928. Modern soybean cultivars present several morphological improvements such as better canopy structure, which enables more efficient light interception and conversion to biomass (Koester et al., 2014).

The first step to develop improve cultivars is to select germplasm with the desired traits and develop a population by crossing two or more parents. Segregation, genetic recombination, and selection will enable the accumulation of favorable alleles and the development of improved progeny (Wilcox, 1998). The study of Quantitative Trait Locus (QTL) can be done simultaneously to breeding to enable the identification of the genomic locations through mapping. The segregating progeny from breeding populations can be used to establish the association between molecular markers and phenotypes and enable the location of genomic regions of relevance.

Within the plant breeding framework, another strategy to improve crop is to generate new genetic variability in a population using artificial mutations through radiation, chemical treatment, or more targeted approaches such as transgenics. These are valuable methods that have proven to be successful to improve traits such as protein (Prenger et al., 2019a), oil composition (Haun et al., 2014) and sucrose (Dobbels et al., 2017).

Soybean Seed Protein

The protein content in soybean seeds is usually 40% and values ranging from 32 to 58 % can be found in the United States Department of Agriculture (USDA) Soybean Germplasm Collection (USDA, 2023b). Traditionally, soybeans have been processed to produce oil and the resulting meal has been a byproduct of the process. As the worldwide demand for animal derived products is increasing, soybean meal became the most important component in livestock feed due to its availability, nutritional properties, and high protein content (Kumar et al., 2010).

Soybean yields in the United States increased 40.8 % in the past 33 years (2241 kg ha^{-1} in 1986 to 3445 kg ha^{-1} in 2021) and in the same period, the protein content went in the opposite direction, decreasing from 35.8 to 33.5% (Naeve & Miller-Garvin, 2021). Growers historically have had no incentive to produce and deliver soybeans with high protein and therefore no focus was given in improving this seed component. This reduction in protein level has negative effects on the soybean value, as lower protein levels imposes difficulty to meet the requirements of the livestock industry to obtain the optimum feed conversion.

The United Soybean Board has defined goals to increase the protein levels of the soybean grown in the United States through the Better Bean Initiative (BBI) to increase the value and competitiveness of the American soybean in the international market (Sallstrom, 2002; Durham, 2003). Increasing the content of protein in seeds would enhance the conversion rate of grains into

meal and reduce the need for supplementation in the livestock feed (Cober & Voldeng, 2000). Some additional benefits can be expected in increasing the seed protein content. Using high protein soybean meal is associated with an increase of milk yield in cows in comparison with meal produced with conventional soybean (McNiven et al., 1994).

A system that compensates farmers for delivering soybeans with higher protein concentration and stimulates the development of new soybean cultivars with improved seed composition is yet to be implemented. One of the obstacles for the establishment of a compensation system is the need for assessing the protein concentration in samples from all the producers delivering to a grain processor (Hurburgh 1994; Maltsbarger and Kalaitzandonakes 2000). Additionally, there is the inherent possibility of mixing high protein with low protein cultivars when the grains are received and processed, unless different systems are used to process the different types of soybeans. These issues and the consequent absence of immediate economic benefits have delayed the development and use of high protein cultivars.

An additional challenge for using high protein genotypes is the economic effects of the correlations with other traits in soybean. It is well known that protein tends to have a negative association with yield and oil and high yield cultivars have the preference of the farmers and the market (Leffel, 1989). This relationship demonstrates that the use of cultivars with improved protein content depends on their ability to have yields at least at the same levels as in the low protein cultivars. It is considered that if the increase in protein content is accompanied by a decrease in oil and yield, the use of the high protein cultivar does not bring economic advantages for the producer (Greiner, 1990).

Genetic control of protein

The heritability of seed protein content is usually high, 0.95 (Chung et al., 2003), 0.93 (Warrington et al., 2015), 0.83 (Lee et al., 1996), 0.66 (Phansak et al., 2016), which may be an indication that genetic factors play a major role in determining the phenotype. In line with this assumption, several QTLs associated with protein content have been identified multiple times in the same region in the genome (Diers et al., 1992; Lee et al., 1996; Brummer et al., 1997;Sebolt et al., 2000; Csanádi et al., 2001; Chung et al., 2003; Bolon et al., 2010; Vaughn et al., 2014 Hwang et al., 2014; Bandillo et al., 2015; Warrington et al., 2015; Patil et al., 2018).

One of the genomic regions frequently reported to be associated with seed protein content is a QTL located on Chr 20 (Table 1.1). The first time this locus was detected in association with high protein was reported by Diers et al. (1992). Using RFLP markers to genotype a population formed by the cross of an elite parent (A81356022) and a *G. soja* line (PI 468916) with high protein content, the authors identified QTLs in four linkage groups (I, E, F, and G) which accounted for 12 to 42% of the phenotypic variation. The protein QTL on Chr 20 (LG-I) stands out because of its high effect on phenotypic variance and its stability, as it has been mapped in an interval between 30 and 34 Mb on Chr. 20 in several studies with different populations tested in multiple environments (Hwang et al., 2014; Vaughn et al., 2014; Warrington et al., 2015). Chung et al. (2003) showed that 65 to 85% of the phenotypic variation for protein and oil could be attributed to the QTL on Chr 20 (LG I). The authors also indicated that the high protein phenotype caused by the Chr 20 QTL was always associated with reduction in oil. This agrees with the hypothesis of a single pleiotropic locus that controls protein and oil content and that these two seed components have an antagonist relationship (Wilcox & Cavins, 1995).

Another important QTL that has been identified multiple times is located on Chr 15 (Diers et al., 1992; Fasoula et al., 2004). Like the Chr 20 QTL, the Chr 15 QTL has been repeatedly mapped in multiple studies (Bandillo et al., 2015; Warrington et al., 2015; Lee et al., 2019). This QTL was mapped to an interval between 3.58 and 4.12 Mb and it was estimated to explain 25.5% of the phenotypic variation and has an additive effect of 0.9% (Kim et al., 2016). Similarly, to the Chr 20 QTL, Chr 15 presented a negative effect on oil content (-0.5%). These two loci are important resources for the improvement of protein content in soybean.

Genetic improvement of protein content

The USDA Soybean Germplasm Collection has approximately 732 accessions with protein content above 50%, which shows that there is genetic variability to provide improvements in this trait (USDA, 2023b). Protein content has a great variation across different accessions and this variation can be used to increase the overall protein content and quality. Older cultivars and plant introductions (PIs) are also an important source of variation that may be employed to improve the seed composition. Vaughn et al. (2014), Bandillo et al. (2015) and Patil et al. (2017) have reported that soybean accessions from South Korea usually present a higher concentration of seed protein than genotypes from the United States and other countries. This is likely a result of the historical breeding efforts in South Korea that focused on the improvement of traits for human consumption especially protein and other functional components (Lee et al., 2015). Accessions such as these from the South Korea are an important source of genetic variability that can be used in breeding programs to ensure a constant improvement of nutritional composition of soybean.

Soybean lines with protein composition reaching values of 47 % have been developed using conventional methods such as recurrent selection and backcross (Wilcox & Cavins, 1995)

and examples of high protein cultivars include Protana (Probst et al., 1971), Prolina (Burton et al., 1999), and Prohio (Mian et al., 2008). However, these cultivars generally show lower yield in comparison with other elite lines with normal protein content. Despite the negative relationship between protein and yield, there are reports on the feasibility of developing lines with increased protein content and yield parity to high performing cultivars (Cober & Voldeng, 2000; Brzostowski et al., 2017).

Relationship between protein content and yield

To develop and effectively deploy high protein cultivars, it is necessary to understand the impacts of the seed protein increase on other traits. Several studies have indicated a negative relationship between protein and yield, with correlation values reaching up to -0.62 (Cober & Voldeng, 2000; Sebolt et al., 2000; Cunicelli et al., 2019). It has been shown that an increase of 1 ton ha^{-1} in yield was associated with a reduction of 2.34 to 2.86% in seed protein (Chung et al., 2003). Overall, a negative correlation between these two traits appears to be common, but contrary to the omnipresent antagonist relationship between oil and protein, in several cases protein and yield do not have a clear association when different populations and different environments are considered.

More recently, Brzostowski et al. (2017) showed that the introgression of the high protein allele from the cultivar Danbaekkong caused a reduction in yield ranging from 273 to 558 kg ha^{-1}. Goettel et al. (2022) indicated that the high protein allele on Chr 20 QTL is associated with a yield decrease of 150.3 kg ha^{-1}. Although specific QTLs might cause this antagonistic relationship between protein and yield, soybean breeders have been able to combine high yield and improved protein content to release new germplasm. Chen et al. (2017) developed UA 5814HP as a new soybean cultivar with high seed protein content (45.5%) and yield comparable

to elite checks. Similarly, Pantalone & Smallwood, (2018) released TN11-5102 as a high yield and high protein line with 42% protein. Prenger et al. (2019b) developed Benning HP, which is an elite line that carries a high-protein allele on Chr 20. This line was developed by backcrossing an F_5-derived line from Benning × Danbaekkong to the recurrent parent Benning. Benning HP exhibited a high protein content (45.9%) and yield equivalent to the recurrent parent Benning in 14 environments over four years (100.3%). These results exemplify the possibility of combining high protein with high yield with progeny selection in breeding populations.

Relationship between protein and oil contents

The negative association between oil and protein is well documented (Cober & Voldeng, 2000; Vaughn et al., 2014; Patil et al., 2018). This relationship is dictated by a ratio of 1:1.7 in which oil and protein compete for the same source of energy and the resources needed to synthesize 1 unit of oil corresponds to the synthesis of 1.7 units of protein (Chung et al., 2003). Although there is a strong negative correlation between protein and oil, there are reports of QTLs that can increase oil without effects on protein. Lee et al. (2019) identified a QTL on chromosome 5 that increases the oil content with no effect on protein content. The authors also identified a QTL on chromosome 10 for protein with low effect on oil.

In an attempt to increase both components, one could try to enhance protein and oil at the expenditure of carbohydrates. It has been demonstrated that the suppression of soybean lipase gene SUGAR-DEPENDENT1 (SDP1) can increase seed oil content and decrease raffinose and stachyose (Aznar-Moreno et al., 2022). The authors showed that suppression of SDP1 through RNAi during seed development increases fatty acid content while reducing oligosaccharides. The authors also demonstrated that there was a small increase in protein content because of additional available carbon that was not used for carbohydrate biosynthesis. SDP1 hydrolyzes stored

triacylglycerols, releasing fatty acids and carbon. Preventing the lipid turnover by SDP1 simultaneously increases oil content and reduces the levels of undesirable carbohydrates (Aznar-Moreno et al., 2022).

Relationship between protein and amino acids contents

The relationship of protein content and the amino acid profile is important for the improvement of soybeans. Researchers have reported that the increase in protein concentration is associated with reduction in protein quality, as the content of essential amino acids such as lysine, threonine, cysteine, and methionine tend to decrease per unit of protein (Panthee et al., 2006a; b). Paek et al. (1997) indicated that when seed protein content is high, gene expression of the sulfur-poor 7S fraction increases in comparison with the 11S. As a result, the methionine and cysteine content decreases. However, this relationship does not seem to be the rule, as there are some reports showing no association between the high protein phenotype and sulfur containing amino acids or in some cases, the increase in protein caused an increase in methionine and cysteine (Edwards et al., 2000; Wilcox & Shibles, 2001).

Several studies have shown that the 7S fraction of protein is negatively associated with the 11S. This relationship seems to be affected by the environment, as sulfur deficiency or the excess of nitrogen in the soil are associated with increase of the 7S fraction and decrease in the 11S (Imsande, 2001; Krishnan, 2005; Wang et al., 2014). This relationship influences the protein quality because the decrease of the 11S fraction reduces the overall concentration of cysteine and methionine in the protein (Krishnan, 2005). QTLs associated with these specific protein fractions have been identified and they can be used to improve the protein quality. QTLs for the protein subunits glycinin (11S) have been identified on Chrs 3, 10, ,13, 17, 20, and 19 and for β-conglycinin (7S) on Chrs 10, 17 and 16 (Panthee et al., 2004; Boehm et al., 2018).

Relationship between protein content, maturity, and environment

The high protein phenotype in some cases is also associated with other traits in addition to seed composition and yield. Simpson & Wilcox (1983) indicated that late maturity genotypes have higher seed protein content. On the other hand, Sebolt et al. (2000) observed that lines with high-protein QTLs from *G. soja* reached maturity earlier. It has been indicated that a locus controlling maturity (E4) is relatively close to the high protein QTL on Chr 20 (Chung et al., 2003; Liu et al., 2008). Patil et al. (2018) also identified QTLs for protein, oil, and sucrose content in *G. max*, co-localized with the maturity loci E1 and E4, and the genetic linkage between loci can create the association between protein and maturity. Patil et al. (2017) reported that cultivars in maturity groups V to X tend to have higher concentration of protein and lower concentration of oil than early maturity cultivars in the maturity groups 000 to II when analyzing data from the USDA Soybean Germplasm. It has also been observed that lines in the MG IV had higher concentration of protein and lower oil content than lines in the MG I (Lee et al., 2019).

Soybean grown in warmer regions with a longer growing season have historically shown higher protein content in comparison with those cultivated in cooler areas (Yaklich et al., 2002). Additionally, drought periods have been associated with a reduction in seed protein content (Specht et al., 2001). Dornbos & Mulle, (1992) indicated that plants grown in a higher temperature (35°C) during the seed filling stage produced seeds with high protein content than those in a lower temperature condition (29°C). In this context, a geographical pattern can be observed across the soybean growing regions in the United States, where the protein content is lower in northern states and it gradually increases towards the southern states (Rotundo et al., 2016). This gradient variation in protein content has also been observed by other researchers (Breene et al., 1988; Yaklich et al., 2002).

It is possible that the maturity effect on protein is also confounded with the environment since late maturity genotypes are planted in the southern region and early maturity in the northern region of the United States. This spatial variation is associated with the differences in temperature, soil and water availability and these environmental factors are considered important in determining the final protein content (Rotundo & Westgate, 2009; Dornbos & Mullen, 1992).

There is also considerable variation in soybean protein composition among the main soybean producing countries. In general, the protein content of soybean produced in the United States is lower than in Brazil, however, the quality of the protein produced in the United States in higher because of a higher percentage of essential amino acids (lysine, cysteine, methionine, threonine, and tryptophan) (Thakur & Hurburgh, 2007). This variation between countries is likely due to differences in environmental conditions, such as temperature and water availability (Patil et al., 2017).

Soybean Seed Amino Acid Composition

Soybean seeds usually have 40% protein, and this makes it an attractive source for the production of meal for livestock. However, in addition to the quantity of protein, the balance of the amino acid composition of the protein is important. The nutritional value of the protein largely depends on the content of essential amino acids.

Soybean meal is primarily used as a protein source for poultry production. The main role of the protein in the diet is to provide a sufficient amount of amino acids for the development of the organism (Friedman & Brandon, 2001). Amino acids are defined as essential or non-essential depending on the metabolism of each animal species. Essential amino acids must be obtained through the diet, as they are not produced in the body. Non-essential amino acids are synthesized by the organism and therefore are not limiting for growth and development. For poultry,

methionine, lysine, threonine, tryptophan, isoleucine, arginine, and valine are essential (Baker, 2003). Most of these amino acids are provided in the soybean meal in sufficient amount with the exception of methionine and cysteine (Pfarr et al., 2018).

It has been demonstrated that an increase in total protein content can cause a relative reduction of five essential amino acids: lysine, cysteine, methionine, threonine, and tryptophan. This is because when estimating the content of amino acids in high protein soybean cultivars by calculating the content based on a seed dry mass, the concentration of amino acids is positively correlated with the protein content of soybeans (Medic et al., 2014). However, when the amino acids content is calculated as percentage of the soybean protein, the percentages of the essential amino acids decreases when protein content is higher. This is known as a dilution effect because the increase in total protein is directed towards non-essential amino acids, such as glutamic acid and arginine, while the content of essential amino acids remains static or in some cases, decreases (Pfarr et al., 2018).

The soybean protein contains all the essential amino acids, however their concentration is low, especially for the two proteogenic sulfur-containing amino acids, methionine and cysteine (Krishnan and Jez, 2018; Pfarr et al., 2018). These amino acids have great importance in the growth and development of animals because methionine is the amino acid responsible for the initiation of protein synthesis and cysteine is crucial for the formation of disulfide bonds (Brosnan & Brosnan, 2006). When poultry is fed on a diet of low sulfur-amino acid content, there is a reduction in weight gain (Conde-Aguilera et al., 2013) and the animals are more susceptible to diseases (Wu, 2014).

The content of these two amino acids together in soybean seeds is approximately 2.6 g per 100 g of protein and this value is lower than the recommended 3.5 g per 100 g of protein

intake (Shewry, 2000). To complement the low concentration of the sulfur-containing amino acids in the meal, livestock producers often add synthetic methionine to the meal to meet the dietary requirements for optimal animal development. These additives represent an increase in the cost of production, which impacts the final price of animal derived products in the market (Imsande, 2001; Krishnan, 2005). To help change this situation, the Better Bean Initiative established the goal of increasing the methionine and cysteine concentrations in the soybean cultivars developed in the United States. The development of new soybean cultivars with improved amino acid content is one of important objectives because it will reduce the costs associated with synthetic amino acids that livestock producers must cover to provide feed for the animals (Durham, 2003). Additionally, the use of soybean meal with high sulfur-containing amino acids is a more sustainable approach because the production of synthetic methionine generates hazardous waste including cyanide, phenols, and benzene and is dependent on fossil fuels (Neubauer & Landecker, 2021).

Although there are reports on the genetic control of soybean amino acid profile (Table 1.2), further studies are important to identify novel genomic regions controlling these components and to develop methods that incorporate this information into breeding programs to generate improved cultivars.

Relationship among different amino acids

In soybean cultivars with high protein, amide amino acids such as asparagine, glutamate, arginine, and glutamine have a higher concentration and amino acids such as threonine, glycine, methionine and histidine tend to have a lower concentration (Serretti et al., 1994; Hernández-Sebastià et al., 2005). Fallen et al. (2013) found positive correlation between methionine and cysteine ($r = 0.76$) and these two amino acids also had positive association with all other amino

acids except for a lysine. Panthee et al. (2006a) reported that methionine and cysteine were positively associated with most of the other amino acids, except with valine. These results indicated that the improvements on cysteine and methionine in soybean can be achieved without drawbacks in the concentration of other essential amino acids.

The levels of individual amino acids in the total soybean protein have been analyzed by de Borja Reis et al. (2020). Glutamic acid and arginine had a positive association with protein, when comparing soybean cultivars released from 1980 to 2014. On the other hand, lysine, valine, proline, alanine, glycine, threonine, tyrosine, and tryptophan had a negative association with protein. Cysteine, and methionine had no significant association with protein. When comparing the relationship between individual amino acids, a positive correlation group was observed among alanine, aspartic acid, glutamic acid, glycine, and serine, with correlation coefficients greater than 0.6. Another group of highly correlated amino acids is isoleucine, leucine, lysine, phenylalanine, threonine, tyrosine, and valine with coefficients greater than 0.48. However, the correlation between amino acids in the different groups tend to be negative (Qin et al., 2019).

Genetic control of individual amino acids

Genomic regions associated with individual amino acid concentration are limited in the literature (Panthee et al., 2006a; b; Fallen et al., 2013; Vaughn et al., 2014; Warrington et al., 2015; Lee et al., 2019) and the only confirmed QTLs for amino acids were identified by Panthee et al. (2006a; b). These authors detected QTLs associated with cysteine on Chrs 1, 13, and 18, and methionine on Chrs 13, 18, and 7. The individual QTLs had and R^2 ranging from 7.6 to 17.6% and each one had a small additive effect, ranging from 0.02 to 0.05 %. In another study comprehending additional amino acids, Panthee et al. (2006b) identified loci associated with lysine concentration on Chrs 1, 15, and 18 and threonine on Chrs 5, 2, 9, and 19. Fallen et al.

(2013) reported three loci on Chr 13 associated with amino acids concentration. Two of these QTLs are very close to previously reported loci associated with protein content, which could be an indicative that in some cases, QTLs for protein content can be associated with protein quality in soybean (Brummer et al., 1997; Reinprecht et al., 2006).

Protein quality improvement

Soybean lines with high protein and increased cysteine concentration have been developed. Four soybean lines (BARC-6, BARC-7, BARC-8, and BARC-9) have a protein content ranging from 49.5 to 53% (Leffel, 1992). In addition to high protein, BARC-8 has a high concentration of cysteine. More recently, a new soybean germplasm with increased concentration of protein and sulfur containing amino acid (methionine + cysteine) was released (Panthee & Pantalone, 2006). TN04-5321 had a protein content of 43.1% and a content of methionine + cysteine of 3.3% based on the average of six environments.

In additional to the conventional strategy of identifying germplasm that carry alleles with positive effects on amino acid composition and the development of populations to introgress the traits into elite germplasm, other approaches are available to improve the nutritional quality of soybeans. Mutation breeding with ethyl methanesulfonate (EMS) has been used to increase the content of specific amino acid and in some cases, soybean lines with concentration of methionine and cysteine 20% higher than the original lines have been obtained (Imsande, 2001).

Transgenic methods have also been employed to change the amino acid composition. The genes encoding the maize 15 kDa zein protein have been transformed into soybeans aiming to increase the methionine and cysteine levels (Dinkins et al., 2001). The transgenic soybean lines containing the maize 15 kDa zein protein gene had a 12 to 20% increase in methionine, and a 15

to 35% increase in cysteine compared to the control. Interestingly, there were no changes in the content of other amino acids in the transgenic lines.

The improvement of amino acid content in soybean has not been the focus in the breeding programs in the recent years and compared to total protein, fewer studies have focused on the identification of the genomic regions associated with individual amino acids. According to Patil et al. (2017), one of the limitations of breeding and studying individual amino acids in soybeans is the narrow genetic variability for the trait and lack of a cost-effective platform for phenotyping seed composition. However, according to Clarke & Wiseman (2000) there is genetic variation in soybean germplasm accessions for amino acid composition. In fact, in the USDA Soybean Germplasm Collection, there are 870 *G. max* accessions in MGs 0 to X that have protein above 43% and Cys + Met above 3%, with values reaching up to 4.7% (USDA, 2023b). This variability can be explored to improve our understanding of the genetic control of the amino acids in the protein.

Technologies to accelerate breeding

Marker-assisted selection

New genomic tools and breeding approaches are important to accelerate the development of new soybean cultivars with improved seed composition. One of the most important tools currently available to accelerate breeding is marker assisted selection (MAS). In this approach the markers identified in mapping and association studies are used to track the traits, reducing the need for phenotyping in each generation. MAS has been used to assist the development of soybean lines with elevated concentrations of the α subunit of β-conglycinin (BC) aiming to improve the quality of soybean protein (Oltmans-Deardorff et al., 2013). Marker Assisted Backcross Breeding (MABB) has been used to reduce the levels of Kunitz trypsin inhibitor

(KTI) in soybean, which are proteins that reduce growth and metabolism of animals. Using markers, the low Kunitz trypsin inhibitor (kti) was introgressed into elite soybean cultivars (Maranna et al., 2016). Another example of the importance of the use of markers to improve seed composition was the development of the high yield, high protein cultivar Benning HP (Prenger et al., 2019). This line is a near isogenic line (NIL) developed by backcrossing a recombinant inbred line from the population Benning × Danbaekkong (Warrington et al., 2015) to the recurrent parent Benning. The backcrossing was performed with MAS using markers flanking the Chr 20 QTL for protein.

In some cases, conventional MAS might not be efficient because some traits are controlled by many loci with small effects. In these cases, marker-assisted recurrent selection (MARS) can be used as an efficient method to combine multiple loci (Varshney et al., 2013). Alternatively, whole-genome prediction can be employed to calculate the effects of all loci associated with a trait and enable the identification of lines for future crosses. It has been demonstrated that genome prediction is able to capture the variation of all loci and accumulate favorable alleles for sulfur-containing amino acids (Miller et al., 2023).

Next generation sequencing

Next-generation sequencing (NGS) is an important tool to understand the evolution of soybean and to enable the discovery of variants associated with the traits of interest. Zhou et al. (2015) sequenced 302 soybean accessions and performed a genome-wide association study and identified two loci associated with oil content co-located with regions of domestication-selective sweeps. More recently, Fliege et al. (2022) and Goettel et al. (2022) compared the genomes *G. max* and *G. soja* accessions and identified an insertion of 321 bp in *Glyma.20G085100*

associated with the Chr 20 QTL for protein, finally resolving this locus after 30 years of research (Diers et al., 1992).

Bayer et al. (2021) analyzed a soybean pangenome with 1,110 soybean accessions derived from the USDA Soybean Germplasm Collection, including 157 *G. soja.* The authors observed a reduction in the number of genes during the transition from *G. soja* to *G. max.* *G. soja* has ~ 600 more genes than *G. max* cultivars and 98 genes with low frequency in *G. max* are associated with defense response. Thirteen of these 98 defense response genes are positioned in known resistance loci, including Sclerotinia resistance, brown stem rot resistance, and Phytophthora resistance. This research exemplifies the importance of large-scale sequencing projects to elucidate the evolution of crops at the genomic level.

The increasing availability of soybean sequencing information opens new opportunities to study genetic variation on a broader scale and enables the validation of discoveries. Recently Zhang et al. (2022) consolidated 1501 soybean genome sequences, including previously published genomes and newly sequenced accessions. The authors provided a comprehensive analysis of the genetic diversity and identification of variants and made the data publicly available. This data set enables the expansion of genetic analyses to a pangenome scale in soybean and further advances the knowledge of the genetic control of traits in this crop.

Induced mutations and transgenic approaches

Improvement in seed composition has become a major objective in soybean breeding programs. Although the natural genetic variation in germplasm collections has served well the selection of lines with improved seed composition, the creation of novel variation can further improve the levels of important components, such as protein, oil and amino acids. Mutagenesis has been an important resource for soybean, and several traits have been modified, such as oil

content (Bolon et al., 2014), stearic acid (Gillman et al., 2014), protein (Prenger et al., 2019a) and sucrose (Dobbels et al., 2017). The collection of mutant lines derived from the soybean fast neutron mutation project is available for the soybean community and it is an important resource for the improvement of this crop (soybase.org/mutants/). Although mutations can be a fast way to induce genetic variation, there are negative effects associated with these methods due to the random nature of the induced change and the consequent effects on non-target traits (Gillman et al., 2014).

In comparison to mutagenesis, genetic engineering methods have a better precision and can improve specific traits without detrimental effects of large random mutations. Precise changes in the soybean genome have been used to improve seed composition traits, such as protein (Schmidt et al., 2011), total oil (Lardizabal et al., 2008), and oleic acid (Haun et al., 2014). This shows that in addition to exploring the natural variation of soybean, the use of transgenic approaches is a valuable tool. In summary, the combination of novel approaches in genomics and molecular genetics expands our knowledge about how traits are controlled and can accelerate breeding of soybeans with improved seed composition.

Objectives

1) Introgress and the Danbaekkong high protein QTL on Chr 20 in a panel of elite soybean lines.
2) Map QTLs controlling sulfur-containing amino acid concentration in soybean and develop breeding lines with improved seed composition.

References

Arnold, B., Menke, E., Mian, M.A.R., Song, Q., Buckley, B., & Li, Z. (2021). Mining QTLs for elevated protein and other major seed composition traits from diverse soybean germplasm.

Molecular Breeding 41(8): 1–18. https://doi.org/10.1007/s11032-021-01242-z.

Aznar-Moreno, J.A., Mukherjee, T., Morley, S.A., Duressa, D., Kambhampati, S., Chu, K.L., Koley, S., Allen, D.K., & Durrett, T.P. (2022). Suppression of SDP1 Improves Soybean Seed Composition by Increasing Oil and Reducing Undigestible Oligosaccharides. *Frontiers in Plant Science* 13(23): 863254. https://doi.org/10.3389/fpls.2022.863254.

Baker, D. (2003). Ideal amino acid patterns for broiler chicks. In: D'Mello, J., editor, Amino Acids in Animal Nutrition. 2nd ed. CABI, Cambridge, MA. p. 223–235

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., & Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *The Plant Genome* 8(3): 1–13. https://doi.org/10.3835/plantgenome2015.04.0024.

Bayer, P.E., Yuan, Y., Batley, J., Nguyen, H.T., Valliyodan, B., Varshney, R.K., Hu, H., Lam, H., Marsh, J.I., Edwards, D., Patil, G., & Song, Q. (2021). Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *The Plant Genome* 15: e20109. https://doi.org/10.1002/tpg2.20109.

Boehm, J.D., Nguyen, V., Tashiro, R., Anderson, D., Chun, S., Wu, X., LWoodrow, L., Yu, K., Cui, Y., & Li, Z. (2018). Genetic mapping and validation of the loci controlling 7S α ′ and 11S A - type storage protein subunits in soybean [*Glycine max* (L .) Merr .]. *Theoretical and Applied Genetics* 131: 659–671. https://doi.org/10.1007/s00122-017-3027-9.

Bolon, Y.T., Joseph, B., Cannon, S.B., Graham, M.A., Diers, B.W., Farmer, A.D., May, G.D., Muehlbauer, G.J., Specht, J.E., Tu, Z.J., Weeks, N., Xu, W.W., Shoemaker, R.C., & Vance, C.P. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biology* 10(41).

https://doi.org/10.1186/1471-2229-10-41.

Bolon, Y.T., Stec, A.O., Michno, J.M., Roessler, J., Bhaskar, P.B., Ries, L., Dobbels, A.A., Campbell, B.W., Young, N.P., Anderson, J.E., Grant, D.M., Orf, J.H., Naeve, S.L., Muehlbauer, G.J., Vance, C.P., & Stupar, R.M. (2014). Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. *Genetics* 198(3): 967–981. https://doi.org/10.1534/genetics.114.170340.

de Borja Reis, A.F., Tamagno, S., Moro Rosso, L.H., Ortez, O.A., Naeve, S., & Ciampitti, I.A. (2020). Historical trend on seed amino acid concentration does not follow protein changes in soybeans. *Scientific Reports* 10(1): 1–10. https://doi.org/10.1038/s41598-020-74734-1.

Breene, W.M., Lin, S., Hardman, L., & Orf, J. (1988). Protein and oil content of soybeans from different geographic locations. *Journal of the American Oil Chemists' Society* 65(12): 1927–1931. https://doi.org/10.1007/BF02546009.

Brosnan, J., & Brosnan, M. (2006). 5th Amino Acid Assessment Workshop. *The Journal of Nutrition* 136(6): 1636–1640.

Brummer, E.C., Graef, G.L., Orf, J., Wilcox, J.R., & Shoemaker, R.C. (1997). Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Science* 37(2): 370–378. https://doi.org/10.2135/cropsci1997.0011183X003700020011x.

Brzostowski, L.F., & Diers, B.W. (2017). Agronomic evaluation of a high protein allele from PI407788A on chromosome 15 across two soybean backgrounds. *Crop Science* 57(6): 2972–2978. https://doi.org/10.2135/cropsci2017.02.0083.

Brzostowski, L.F., Pruski, T.I., Specht, J.E., & Diers, B.W. (2017). Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. *Theoretical and Applied Genetics* 130(11): 2315–2326. https://doi.org/10.1007/s00122-017-2961-x.

Burton, J.W., Carter, T.E., & Wilson, R.F. (1999). Registration of ` Prolina ' Soybean. *Crop Science* 39(1): 1993–1994. https://doi.org/10.2135/cropsci1999.0011183X003900010066x.

Chen, P., Florez-Palacios, L., Orazaly, M., Manjarrez-Sandoval, P., Wu, C., Rupe, J.C., Dombek, D.G., Kirkpatrick, T., & Robbins, R.T. (2017). Registration of 'UA 5814HP' Soybean with High Yield and High Seed-Protein Content. *Journal of Plant Registrations* 11(2): 116–120. https://doi.org/10.3198/jpr2016.09.0046crc.

Chung, J., Babka, H.L., Graef, G.L., Staswick, P.E., Lee, D.J., Cregan, P.B., Shoemaker, R.C., & Specht, J.E. (2003). The Seed Protein, Oil, and Yield QTL on Soybean Linkage Group I. *Crop Science* 43: 1053–1067. https://doi.org/10.2135/cropsci2003.1053.

Clarke, E., & Wiseman, J. (2000). Developments in plant breeding for improved nutritional quality of soya beans I. Protein and amino acid content. *The Journal of Agricultural Science* 134(2): 111–124.

Cober, E.R., & Voldeng, H.D. (2000). Developing high-protein, high-yield soybean populations and lines. *Crop Science* 40(1): 39–42. https://doi.org/10.2135/cropsci2000.40139x.

Conde-Aguilera, J.A., Cobo-Ortega, C., Tesseraud, S., Lessire, M., Mercier, Y., & van Milgen, J. (2013). Changes in body composition in broilers by a sulfur amino acid deficiency during growth. *Poultry Science* 92(5): 1266–1275. https://doi.org/https://doi.org/10.3382/ps.2012-02796.

Csanádi, G., Vollmann, J., Stift, G., & Lelley, T. (2001). Seed quality QTLs identified in a molecular map of early maturing soybean. *Theoretical and Applied Genetics* 103(6–7): 912–919. https://doi.org/10.1007/s001220100621.

Cunicelli, M.J., Bhandari, H.S., Chen, P., Sams, C.E., Mian, M.A.R., Mozzoni, L.A., Smallwood, C.J., & Pantalone, V.R. (2019). Effect of a Mutant Danbaekkong Allele on

Soybean Seed Yield, Protein, and Oil Concentration. *Journal of the American Oil Chemists' Society* 96: 927–935. https://doi.org/10.1002/aocs.12261.

Diers, B.W., Keim, P., Fehr, W.R., & Shoemaker, R.C. (1992). RFLP analysis of soybean seed protein and oil content. *Theoretical and Applied Genetics* 83(5): 608–612. https://doi.org/10.1007/BF00226905.

Dinkins, R.D., Srinivasa Reddy, M.S., Meurer, C.A., Yan, B., Trick, H., Thibaud-Nissen, F., Finer, J.J., Parrott, W.A., & Collins, G.B. (2001). Increased sulfur amino acids in soybean plants overexpressing the maize 15 kDa zein protein. *In Vitro Cellular and Developmental Biology* 37(6): 742–747. https://doi.org/10.1007/s11627-001-0123-x.

Dobbels, A.A., Michno, J., Campbell, B.W., Virdi, K.S., Stec, A.O., Muehlbauer, G.J., Naeve, S.L., & Stupar, R.M. (2017). An Induced Chromosomal Translocation in Soybean Disrupts a KASI Ortholog and Is Associated with a High-Sucrose and Low-Oil Seed Phenotype. *G3 Genes Genomes Genetics* 7(4): 1215–1223. https://doi.org/10.1534/g3.116.038596/-/DC1.1.

Dornbos, D.L., & Mullen, R.E. (1992). Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. *Journal of the American Oil Chemists Society* 69(3): 228–231. https://doi.org/10.1007/BF02635891.

Durham, D. (2003). The United Soybean Board's better bean initiative: Building United States soybean competitiveness from the inside out. *AgBio Forum* 6(1): 23–26.

Edwards, H.M., Douglas, M.W., Parsons, C.M., & Baker, D.H. (2000). Protein and energy evaluation of soybean meals processed from genetically modified high-protein soybeans. *Poultry Science* 79(4): 525–527. https://doi.org/10.1093/ps/79.4.525.

Fallen, B.D., Hatcher, C.N., Allen, F.L., Kopsel, D.A., Saxton, A.M., Chen, P., Kantartzi, S.K., Cregan, P.B., Hyten, D.L., & Pantalone, V.R. (2013). Soybean Seed Amino Acid Content

QTL Detected Using the Universal Soy Linkage Panel 1.0 with 1,536 SNPs. *Journal of Plant Genome Sciences* 1(3): 68–79. https://doi.org/10.5147/jpgs.2013.0089.

FAO. (2023). FAOSTAT - Statistics Database. http://www.fao.org/faostat/en/#home (accessed 5 March 2023).

Fasoula, V.A., Harris, D.K., & Boerma, H.R. (2004). Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. *Crop Science* 44(4): 1218–1225.

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., Hudson, M.E., & Diers, B.W. (2022a). Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. *Plant Journal* 110(1): 114–128. https://doi.org/10.1111/tpj.15658.

Friedman, M., & Brandon, D.L. (2001). Nutritional and Health Benefits of Soy Proteins. *Journal of Agricultural and Food Chemistry* 49(3): 1069–1086. https://doi.org/10.1021/jf0009246.

Gillman, J.D., Stacey, M.G., Cui, Y., Berg, H.R., & Stacey, G. (2014). Deletions of the SACPD-C locus elevate seed stearic acid levels but also result in fatty acid and morphological alterations in nitrogen fixing nodules. *BMC Plant Biology* 14(1). https://doi.org/10.1186/1471-2229-14-143.

Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., Song, Q., Pantalone, V.R., Song, B.-H., Yu, D., & An, Y.C. (2022). POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nature Communications* 13: 3051. https://doi.org/10.1038/s41467-022-30314-7.

Greiner, C.A. (1990). Special Report 92: Economic Implications of Modified Soybean Traits, Iowa State University Press, Ames.

Haun, W., Coffman, A., Clasen, B.M., Demorest, Z.L., Lowy, A., Ray, E., Retterath, A., Stoddard, T., Juillerat, A., Cedrone, F., Mathis, L., Voytas, D.F., & Zhang, F. (2014). Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant Biotechnology Journal* 12(7): 934–940. https://doi.org/10.1111/pbi.12201.

Hernández-Sebastià, C., Marsolais, F., Saravitz, C., Israel, D., Dewey, R.E., & Huber, S.C. (2005). Free amino acid profiles suggest a possible role for asparagine in the control of storage-product accumulation in developing seeds of low- and high-protein soybean lines. *Journal of Experimental Botany* 56(417): 1951–1963. https://doi.org/10.1093/jxb/eri191.

Hurburgh, C. R. (1994). Identification and segregation of high-value soybeans at a country elevator. *Journal of the American Oil Chemists' Society* 71: 1073–1078.

Hwang, E.Y., Song, Q., Jia, G., Specht, J.E., Hyten, D.L., Costa, J., & Cregan, P.B. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15(1): 1–12. https://doi.org/10.1186/1471-2164-15-1.

Hymowitz, T., & Harlan, J.R. (1983). Introduction of soybean to North America by Samuel Bowen in 1765. *Economic Botany* 37(4): 371–379. https://doi.org/10.1007/BF02904196.

Hymowitz, T., & Newell, C.A. (1981). Taxonomy of the Genus Glycine , Domestication and Uses of Soybeans. *Economic Botany* 35(3): 272–288.

Hyten, D.L., Pantalone, V.R., Sams, C.E., Saxton, A.M., Landau-Ellis, D., Stefaniak, T.R., & Schmidt, M.E. (2004). Seed quality QTL in a prominent soybean population. *Theoretical and Applied Genetics* 109(3): 552–561. https://doi.org/10.1007/s00122-004-1661-5.

Hyten, D.L., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., & Cregan, P.B. (2006). Impact of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences of the United States of America* 103(45):

16666–16671. https://doi.org/10.1073/pnas.0604379103.

Imsande, J. (2001). Selection of Soybean Mutants with Increased Concentrations of Seed Methionine and Cysteine. *Crop Science* 41: 510–515. https://doi.org/10.2135/cropsci2001.412510x.

Kim, M.S., Lozano, R., Kim, J.H., Bae, D.N., Kim, S.T., Park, J.H., Choi, M.S., Kim, J., Ok, H.C., Park, S.K., Gore, M.A., Moon, J.K., & Jeong, S.C. (2021). The patterns of deleterious mutations during the domestication of soybean. *Nature Communications* 12(1): 1–14. https://doi.org/10.1038/s41467-020-20337-3.

Kim, M., Schultz, S., Nelson, R.L., & Diers, B.W. (2016). Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. *Crop Science* 56(1): 219–225. https://doi.org/10.2135/cropsci2015.06.0340.

Koester, R.P., Skoneczka, J.A., Cary, T.R., Diers, B.W., & Ainsworth, E.A. (2014). Historical gains in soybean (*Glycine max* Merr.) seed yield are driven by linear increases in light interception, energy conversion, and partitioning efficiencies. *Journal of Experimental Botany* 65(12): 3311–3321. https://doi.org/10.1093/jxb/eru187.

Krishnan, H.B. (2005). Engineering Soybean for Enhanced Sulfur Amino Acid Content. *Crop Science* 45: 454–461. https://doi.org/10.2135/cropsci2005.0454.

Kumar, V., Rani, A., & Chauhan, G.S. (2010). Nutritional value of soybean. The Soybean: botany, production and uses. 1ª. CABI. p. 375–403

Lardizabal, K., Effertz, R., Levering, C., Mai, J., Pedroso, M.C., Jury, T., Aasen, E., Gruys, K., & Bennett, K. (2008). Expression of Umbelopsis ramanniana DGAT2A in seed increases oil in soybean. *Plant Physiology* 148(1): 89–96. https://doi.org/10.1104/pp.108.123042.

Lee, S.H., Bailey, M.A., Mian, M.A.R., Carter, T.E., Shipe, E.R., Ashley, D.A., Parrott, W.A.,

Hussey, R.S., & Boerma, H.R. (1996). RFLP loci associated with soybean seed protein and oil content across populations and locations. *Theoretical and Applied Genetics* 93(5–6): 649–657. https://doi.org/10.1007/BF00224058.

Lee, C., Choi, M., Kim, H., Yun, H., Lee, B., Chung, Y., Kim, R., & Choi, H. (2015). Soybean [*Glycine max* (L.) Merrill]: Importance as A Crop and Pedigree Reconstruction of Korean Varieties. *Plant Breeding and Biotechnology* 3(3): 179–196. https://doi.org/10.1016/S0828-282X(08)70684-6.

Lee, S., Van, K., Sung, M., Nelson, R., LaMantia, J., McHale, L.K., & Mian, M.A.R. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theoretical and Applied Genetics* 132(6): 1639–1659. https://doi.org/10.1007/s00122-019-03304-5.

Leffel, R.C. (1989). Breeding Soybeans for the Economic Values of Seed Oil and Protein. *Journal of Production Agriculture* 2: 338–343. https://doi.org/10.2134/jpa1989.0338.

Leffel, R.C. (1992). Registration of High-Protein Soybean Germplasm Lines BARC-6, BARC-7, BARC-8, and BARC-9. *Crop Science* 32(2): 502–502. https://doi.org/10.2135/cropsci1992.0011183x003200020054x.

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., & Abe, J. (2008). Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics* 180(2): 995–1007. https://doi.org/10.1534/genetics.108.092742.

Lu, W., Wen, Z., Li, H., Yuan, D., Li, J., Zhang, H., Huang, Z., Cui, S., & Du, W. (2013). Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. *Theoretical and Applied Genetics* 126(2): 425–433. https://doi.org/10.1007/s00122-012-1990-8.

Magwene, P.M., Willis, J.H., & Kelly, J.K. (2011). The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology* 7(11): 1–9. https://doi.org/10.1371/journal.pcbi.1002255.

Maltsbarger, R., and N. Kalaitzandonakes (2000). Direct and hidden costs in identity preserved supply chains. *AgBioForum* 3: 236–242.

Maranna, S., Verma, K., Talukdar, A., Lal, S.K., Kumar, A., & Mukherjee, K. (2016). Introgression of null allele of Kunitz trypsin inhibitor through marker-assisted backcross breeding in soybean (*Glycine max* L. Merr.). *BMC Genetics* 17(1): 1–9. https://doi.org/10.1186/s12863-016-0413-2.

Marsh, J.I., Hu, H., Petereit, J., Bayer, P.E., Valliyodan, B., Batley, J., Nguyen, H.T., & Edwards, D. (2022). Haplotype mapping uncovers unexplored variation in wild and domesticated soybean at the major protein locus cqProt-003. *Theoretical and Applied Genetics* 135: 1443–1455. https://doi.org/10.1007/s00122-022-04045-8.

https://doi.org/10.1071/AR9800951.

McNiven, M.A., Robinson, P.H., & MacLeod, J.A. (1994). Evaluation of a New High Protein Variety of Soybeans as a Source of Protein and Energy for Dairy Cows. *Journal of Dairy Science* 77(9): 2605–2613. https://doi.org/10.3168/jds.S0022-0302(94)77201-5.

Medic, J., Atkinson, C., & Hurburgh, C.R. (2014). Current knowledge in soybean composition. *JAOCS, Journal of the American Oil Chemists' Society* 91(3): 363–384. https://doi.org/10.1007/s11746-013-2407-9.

Mian, M.A.R., Cooper, R.L., & Dorrance, A.E. (2008). Registration of 'Prohio' Soybean. *Journal of Plant Registrations* 2(3): 208–210. https://doi.org/10.3198/jpr2007.09.0531crc.

Miller, M.J., Song, Q., & Li, Z. (2023). Genomic Selection of Soybean (*Glycine max*) for

Genetic Improvement of Yield and Seed Composition in a Breeding Context. *Plant Genome*. https://doi.org/https://doi.org: 10.1002/tpg2.20384.

Naeve, S., & Miller-Garvin, J. (2021). United States soybean quality - Annual Report, Dep. of Agronomy, University of Minnesota, St. Paul.

Neubauer, C., & Landecker, H. (2021). Personal View A planetary health perspective on synthetic methionine. *The Lancet Planetary Health* 5(8): e560–e569. https://doi.org/10.1016/S2542-5196(21)00138-8.

Nichols, D.M., Glover, K.D., Carlson, S.R., Specht, J.E., & Diers, B.W. (2006). Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits Research supported in part by the Illinois Soybean Program Operating Board and the North Central Soybean Research Program. *Crop Science* 46: 834–839. https://doi.org/10.2135/cropsci2005.05-0168.

Oltmans-Deardorff, S.E., Fehr, W.R., & Shoemaker, R.C. (2013). Marker-assisted selection for elevated concentrations of the α' subunit of β-conglycinin and its influence on agronomic and seed traits of soybean. *Crop Science* 53(1): 1–8. https://doi.org/10.2135/cropsci2012.03.0205.

Paek, N.C., Imsande, J., Shoemaker, R.C., & Shibles, R. (1997). Nutritional Control of Soybean Seed Storage Protein. *Crop Science* 37: 498–503. https://doi.org/10.2135/cropsci1997.0011183X003700020031x.

Pantalone, V., & Smallwood, C. (2018). Registration of 'TN11-5102' Soybean Cultivar with High Yield and High Protein Meal. *Journal of Plant Registrations* 12(3): 304–308. https://doi.org/https://doi.org/10.3198/jpr2017.10.0074crc.

Panthee, D.R., Kwanyuen, P., Sams, C.E., West, D.R., Saxton, A.M., & Pantalone, V.R. (2004).

Quantitative trait loci for β-conglycinin (7S) and glycinin (11S) fractions of soybean storage protein. *Journal of the American Oil Chemists' Society* 81(11): 1005–1012. https://doi.org/10.1007/s11746-004-1014-4.

Panthee, D.R., & Pantalone, V.R. (2006). Registration of Soybean Germplasm Lines TN03–350 and TN04–5321 with Improved Protein Concentration and Quality. *Crop Science* 46(5): 2328–2329. https://doi.org/10.2135/cropsci2005.11.0437.

Panthee, D.R., Pantalone, V.R., Sams, C.E., Saxton, A.M., West, D.R., Orf, J.H., & Killam, A.S. (2006a). Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. *Theoretical and Applied Genetics* 112(3): 546–553. https://doi.org/10.1007/s00122-005-0161-6.

Panthee, D.R., Pantalone, V.R., Saxton, A.M., West, D.R., & Sams, C.E. (2006b). Genomic regions associated with amino acid composition in soybean. *Molecular Breeding* 17(1): 79–89. https://doi.org/10.1007/s11032-005-2519-5.

Panthee, D.R., Pantalone, V.R., West, D.R., Saxton, A.M., & Sams, C.E. (2005). Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. *Crop Science* 45(5): 2015–2022. https://doi.org/10.2135/cropsci2004.0720.

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., Nguyen, H.T., & Sleper, D.A. (2013). Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. *Crop Science* 53(3): 765–774. https://doi.org/10.2135/cropsci2012.03.0153.

Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., Shannon, G.J., Carter, T.C., & Nguyen, H.T. (2017). Molecular mapping and genomics of soybean seed protein: a review

and perspective for the future. *Theoretical and Applied Genetics* 130(10): 1975–1991. https://doi.org/10.1007/s00122-017-2955-8.

Patil, G., Vuong, T.D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C., Wu, X., Bai, Y., Yungbluth, D., Lu, F., Kumpatla, S., Shannon, J.G., Varshney, R.K., & Nguyen, H.T. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. *Plant Biotechnology Journal* 16(11): 1939–1953. https://doi.org/10.1111/pbi.12929.

Pfarr, M.D., Kazula, M.J., Miller-Garvin, J.E., & Naeve, S.L. (2018). Amino acid balance is affected by protein concentration in soybean. *Crop Science* 58(5): 2050–2062. https://doi.org/10.2135/cropsci2017.11.0703.

Phansak, P., Soonsuwon, W., Hyten, D.L., Song, Q., Cregan, P.B., Graef, G.L., & Specht, J.E. (2016). Multi-population selective genotyping to identify soybean [*Glycine max* (L.) Merr.] seed protein and oil QTLs. *G3: Genes, Genomes, Genetics* 6(6): 1635–1648. https://doi.org/10.1534/g3.116.027656.

Prenger, E.M., Ostezan, A., Mian, M.A.R., Stupar, R.M., Glenn, T., & Li, Z. (2019a). Identification and characterization of a fast-neutron-induced mutant with elevated seed protein content in soybean. *Theoretical and Applied Genetics* 132(11): 2965–2983. https://doi.org/10.1007/s00122-019-03399-w.

Prenger, E.M., Yates, J., Mian, M.A.R., Buckley, B., Boerma, H.R., & Li, Z. (2019b). Introgression of a high protein allele into an elite soybean cultivar results in a high-protein near-isogenic line with yield parity. *Crop Science* 59(6): 2498–2508. https://doi.org/10.2135/cropsci2018.12.0767.

Probst, A.H., Laviolette, F.A., Athow, K.L., & Wilcox, J.R. (1971). Registration of Protana

Soybean. *Crop Science* 11(2): 312–312. https://doi.org/10.2135/cropsci1971.0011183x001100020050x.

Qi, Z., Sun, Y., Wu, Q., Liu, C., Hu, G., & Chen, Q. (2011). A meta-analysis of seed protein concentration QTL in soybean. *Canadian Journal of Plant Science* 91(1): 221–230. https://doi.org/10.4141/CJPS09193.

Qin, J., Shi, A., Song, Q., Li, S., Wang, F., Cao, Y., Ravelombola, W., Song, Q., Yang, C., & Zhang, M. (2019). Genome Wide Association Study and Genomic Selection of Amino Acid Concentrations in Soybean Seeds. *Frontiers in Plant Science* 10(11): 1–15. https://doi.org/10.3389/fpls.2019.01445.

Qiu, L.J., & Chang, R.Z. (2010). The origin and history of soybean. In: Singh, G., editor, The Soybean: botany, production and uses. p. 1–23

Reinprecht, Y., Poysa, V.W., Yu, K., Rajcan, I., Ablett, G.R., & Pauls, K.P. (2006). Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. *Genome* 49(12): 1510–1527. https://doi.org/10.1139/g06-112.

Rotundo, J.L., Miller-Garvin, J.E., & Naeve, S.L. (2016). Regional and Temporal Variation in Soybean Seed Protein and Oil across the United States. *Crop Science* 56: 797–808. https://doi.org/10.2135/cropsci2015.06.0394.

Rotundo, L., & Westgate, M.E. (2009). Field Crops Research Meta-analysis of environmental effects on soybean seed composition. *Field Crops Research* 110: 147–156. https://doi.org/10.1016/j.fcr.2008.07.012.

Sallstrom, J.R. (2002). Better Bean Initiative (BBI) - A tool to enhance competitiveness for the U.S. soybean producer. The 9th Bienn. Conf. Cell. Mol. Biol. Soybean. Urbana-Champaign

Schmidt, M.A., Barbazuk, W.B., Sandford, M., May, G., Song, Z., Zhou, W., Nikolau, B.J., &

Herman, E.M. (2011). Silencing of soybean seed storage proteins results in a rebalanced protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. *Plant Physiology* 156(1): 330–345. https://doi.org/10.1104/pp.111.173807.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278): 178–183. https://doi.org/10.1038/nature08670.

Sebolt, A.M., Shoemaker, R.C., & Diers, B.W. (2000). Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Science* 40(5): 1438–1444. https://doi.org/10.2135/cropsci2000.4051438x.

Sedivy, E.J., Wu, F., & Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytologist* 214(2): 539–553. https://doi.org/10.1111/nph.14418.

Serretti, C., Schapaugh, W.T., & Leffel, R.C. (1994). Amino Acid Profile of High Seed Protein Soybean. *Crop Science* 34: 207–209. https://doi.org/10.2135/cropsci1994.0011183X003400010037x.

Shewry, P.R. (2000). Seed proteins. Seed Technology and its Biological Basis. Sheffield Academic Press, Sheffield. p. 42–84

Simpson, A.M., & Wilcox, J.R. (1983). Genetic and Phenotypic Associations of Agronomic Characteristics in Four High Protein Soybean Populations. *Crop Science* 23: 1077–1081. https://doi.org/10.2135/cropsci1983.0011183X002300060013x.

Singer, W.M., Shea, Z., Yu, D., Huang, H., Mian, M.A.R., Shang, C., Rosso, M.L., Song, Q.J., & Zhang, B. (2022). Genome-Wide Association Study and Genomic Selection for

Proteinogenic Methionine in Soybean Seeds. *Frontiers in Plant Science* 13: 859109. https://doi.org/10.3389/fpls.2022.859109.

Singh, R.J., & Hymowitz, T. (1988). The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. *Theoretical and Applied Genetics* 76(5): 705–711. https://doi.org/10.1007/BF00303516.

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I., & Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnology Journal* 13(2): 211–221. https://doi.org/10.1111/pbi.12249.

Specht, J.E., Chase, K., Macrander, M., Graef, G.L., Chung, J., Markwell, J.P., Germann, M., Orf, J.H., & Lark, K.G. (2001). Soybean Response to Water: A QTL Analysis of Drought Tolerance. *Crop Science* 41: 493–509. https://doi.org/10.2135/cropsci2001.412493x.

Thakur, M., & Hurburgh, C.R. (2007). Quality of US Soybean Meal Compared to the Quality of Soybean Meal from Other Origins. *Journal of the American Oil Chemists' Society* 84(9): 835–843. https://doi.org/https://doi.org/10.1007/s11746-007-1107-8.

USDA. (2023a). World Supply and Use of Oilseeds and Oilseed Products. *Oil Crop Yearbook*. https://www.ers.usda.gov/data-products/oil-crops-yearbook/.

USDA. (2023b). Germplasm Resources Information Network (GRIN) - National Plant Germplasm System. https://www.ars-grin.gov/.

Valliyodan, B., Dan, Q., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., Song, L., Vuong, T.D., Musket, T.A., Xu, D., Shannon, J.G., Shifeng, C., Liu, X., & Nguyen, H.T. (2016). Landscape of genomic diversity and trait discovery in soybean. *Scientific Reports* 6: 23598. https://doi.org/10.1038/srep23598\rsrep23598 [pii].

Van, K., & McHale, L.K. (2017). Meta-Analyses of QTLs associated with protein and oil contents and compositions in soybean [*Glycine max* (L.) Merr.] Seed. *International Journal of Molecular Sciences* 18(6). https://doi.org/10.3390/ijms18061180.

Varshney, R.K., Mohan, S.M., Gaur, P.M., Gangarao, N.V.P.R., Pandey, M.K., et al. (2013). Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnology Advances* 31(8): 1120–1134. https://doi.org/https://doi.org/10.1016/j.biotechadv.2013.01.001.

Vaughn, J.N., Nelson, R.L., Song, Q., Cregan, P.B., & Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3: Genes, Genomes, Genetics* 4(11): 2283–2294. https://doi.org/10.1534/g3.114.013433.

Wang, J., Chen, P., Wang, D., Shanon, G., Zeng, A., Orazaly, M., & Wu, C. (2015). Identification and mapping of stable QTL for protein content in soybean seeds. *Molecular Breeding* 35(92). https://doi.org/10.1007/s11032-015-0285-6.

Wang, J., Liu, L., Guo, Y., Wang, Y. hui, Zhang, L., Jin, L. guo, Guan, R. xia, Liu, Z. xiong, Wang, L. lin, Chang, R. zhen, & Qiu, L. juan. (2014). A Dominant Locus, qBSC-1, Controls β Subunit Content of Seed Storage Protein in Soybean (*Glycine max* (L.) Merri.). *Journal of Integrative Agriculture* 13(9): 1854–1864. https://doi.org/10.1016/S2095-3119(13)60579-1.

Wang, J., Mao, L., Zeng, Z., Yu, X., Lian, J., Feng, J., Yang, W., An, J., Wu, H., Zhang, M., & Liu, L. (2021). Genetic mapping high protein content QTL from soybean 'Nanxiadou 25' and candidate gene analysis. *BMC Plant Biology* 21(1): 1–13. https://doi.org/10.1186/s12870-021-03176-2.

Warrington, C. V., Abdel-Haleem, H., Hyten, D.L., Cregan, P.B., Orf, J.H., Killam, A.S., Bajjalieh, N., Li, Z., & Boerma, H.R. (2015). QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. *Theoretical and Applied Genetics* 128(5): 839–850. https://doi.org/10.1007/s00122-015-2474-4.

Wilcox, J.R. (1998). Increasing Seed Protein in Soybean with Eight Cycles of Recurrent Selection. *Crop Science* 38(6): 1536–1540. https://doi.org/https://doi.org/10.2135/cropsci1998.0011183X003800060021x.

Wilcox, J.R., & Cavins, J.F. (1995). Backcrossing High Seed Protein to a Soybean Cultivar. *Crop Science* 35: 1036–1041. https://doi.org/10.2135/cropsci1995.0011183X003500040019x.

Wilcox, J., & Shibles, R. (2001). Interrelationships among Seed Quality Attributes in Soybean. *Crop Science* 41. https://doi.org/10.2135/cropsci2001.41111x.

Wu, G. (2014). Dietary requirements of synthesizable amino acids by animals: a paradigm shift in protein nutrition. *Journal of Animal Science and Biotechnology* 5(1): 34. https://doi.org/10.1186/2049-1891-5-34.

Yaklich, R.W., Vinyard, B., Camp, M., & Douglass, S. (2002). Analysis of Seed Protein and Oil from Soybean Northern and Southern Region Uniform Tests. *Crop Science* 42: 1504–1515. https://doi.org/10.2135/cropsci2002.1504.

Zhang, H., Jiang, H., Hu, Z., Song, Q., & An, Y. qiang C. (2022). Development of a versatile resource for post-genomic research through consolidating and characterizing 1500 diverse wild and cultivated soybean genomes. *BMC Genomics* 23(1): 1–13. https://doi.org/10.1186/s12864-022-08326-w.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., et al. (2015). Resequencing 302 wild and

cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* 33(4): 408–414. https://doi.org/10.1038/nbt.3096.

Table 1.1. Major protein QTLs reported in soybean.

Method	Female Parent	Male Parent	Population size	Markers	Chromosome	References
QTL mapping	A81356022	PI 468916†	60	RFLP (243), isozyme (5), storage protein (1), morphological (3)	20, 15, 18	Diers et al. (1992)
	Young	PI 416937†	120	RFLP (155)	15	Lee et al. (1996)
	M82806	HHP†	71	RFLP (75)	20, 15, 18	Brummer et al. (1997)
	A3733	PI 437088A†	76	SSR (103), RAPD (329)	20	Chung et al. (2003)
	Essex†	Williams	131	SSR (100)	6	Hyten et al. (2004)
	N87-984-16†	TN93-99	101	SSR (94)	18	Panthee et al. (2005)
	ZDD09454†	Yudou12	212	SSR (301)	20, 18	Lu et al. (2013)
	Magellan	PI 438489B†	216	SSR, SNP (900)	15, 5,6	Pathan et al. (2013)
	R05-638	R05-1415†	242	SSR (120), SNP (526)	14, 20	Wang et al. (2015)
	Benning	Danbaekkong†	140	SSR (98), SNP (323)	20, 15	Warrington et al. (2015)
	48 F2 populations		~224 each	SNP (1536)	20, 15,10	Phansak et al. (2016)
	William 82	PI 483460B†	188	WGS (0.3x), SNP (3K)	6,8,13,19,20	Patil et al. (2018)
	Tongdou 11	Nanxiadou 25†	178	16,546 SNPs	All but 4, 12, 14, 17, 18, and 19	Wang et al. (2021)
QTL Validation	Parker	PI 468916†	100	SSR (2), RFLP (3)	20	Sebolt et al. (2000)
	PI 97100†	Coker 237	176	RFLP (4)	15	Fasoula et al. (2004)
Fine mapping	A81356022	PI 468916†	247	SSR (9), AFLP (30)	20	Nichols et al. (2006)
	A81356022	PI 468916†	Multiple	SSR and SNPs	20	Fliege et al., (2022)
	Williams 82	PI 479752†	300		20	Goettel et al. (2022)
Meta-QTL analysis					20,5,7,14,15	Qi et al. (2011)
					20	Van and McHale (2017)
GWAS			298	SoySNP50K	20	Hwang et al. (2014)
			3258	SoySNP50K	20	Vaughn et al. (2014)

			139	GBS-47K	20, 5, 8	Sonah et al. (2015)
			302	WGS	13,3, 17,12,11,15	Zhou et al. (2015)
			12116	SoySNP50K	20,15,6	Bandillo et al. (2015)
			106	WGS	20	Valliyodan et al. (2016)
			621	SoySNP50K	15, 20	Lee et al. (2019)
			781	WGS	5, 8, 13, 15, and 20	Kim et al., (2021)
			985	WGS	20	Marsh et al., (2022)
			278	WGS	20	Goettel et al. (2022)
Transcriptome analysis	A81356022	PI 468916†			20	Bolon et al. (2010)
Effect of high protein QTL	AR09-192019, LD02-4485	LG05C-1782 (PI407788A allele)†			15	Brzostowski et al. (2017a)
	Dwight, LD02-5025, Loda, LS93-0375	PI 468916†, Danbaekkong†			20	Brzostowski et al. (2017b)
	LD00-2817P	G03-3101 (Danbaekkong allele)†			20	Cunicelli et al. (2019)
	Benning	Danbaekkong†			20	Prenger et al. (2019)

Note: RFLP refers to Restriction Fragment Length Polymorphism, SSR are Simple Sequence Repeats, RAPD is Random Amplified Polymorphic DNA, SNP is Single Nucleotide Polymorphism, WGS is Whole-Genome Sequencing, AFLP is Amplified Fragment Length Polymorphism, SoySNP50K refers to Illumina Infinium BeadChip containing 50,000 SNPs and GBS is Genotyping by Sequencing.

† High protein parent.

Table 1.2. QTLs reported in the literature for elevated cysteine and methionine contents.

Trait	Study	Female Parent	Male Parent	Population size	Markers	Chromosome	References
Cysteine	QTL mapping	N87-984-16†	TN93-99	101	94 SSR	1,13,18	Panthee et al. (2006a)
	QTL mapping	Essex†	Williams 82	282	1539 SNPs	20	Fallen et al. (2013)
	QTL mapping	Benning	Danbaekkong†	140	SSR (98), SNP (323)	10	Warrington et al. (2015)
	GWAS			877	SoySNP50K	3	Lee et al. (2019)
	QTL mapping	G00-3213	PI 594458A†	132	SoySNP6K	1,2,10,17,20	Arnold et al. (2021)
Methionine	QTL mapping	N87-984-16†	TN93-99	101	94 SSR	13, 18, 7	Panthee et al. (2006a)
	QTL mapping	Essex	Williams 82	282	1539 SNPs	13	Fallen et al. (2013)
	GWAS			3258	SoySNP50K	1,3,11,16, 20	Vaughn et al. (2014)
	QTL mapping	Benning	Danbaekkong†	140	SSR (98), SNP (323)	6,9,10,20	Warrington et al. (2015)
	GWAS			877	SoySNP50K	1, 15, 18	Lee et al. (2019)
	QTL mapping	G00-3213	PI 594458A†	132	SoySNP6K	6,10,20	Arnold et al. (2021)
	GWAS			311	SoySNP50K	3,4,5,6,8,12,16	Singer et al. (2022)

Note: SSR corresponds to Simple Sequence Repeats, SNP is Single Nucleotide Polymorphism, SoySNP6K refers to Illumina Infinium BeadChip containing 6,000 SNPs, SoySNP50K refers to Illumina Infinium BeadChip containing 50,000 SNPs.
† High protein and high amino acid parent.

CHAPTER 2

INTROGRESSION OF A DANBAEKKONG HIGH PROTEIN ALLELE ACROSS DIFFERENT GENETIC BACKGROUNDS IN SOYBEAN [1]

1 Renan Souza, M. A. Rouf Mian, Justin N. Vaughn, and Zenglu Li. To be submitted to *Frontiers in Plant Science*.

Abstract

Soybean meal is a major component of livestock feed due to its high content and quality of protein. Understanding the genetic control of protein is essential to develop new cultivars with improved meal protein. Previously, a genomic region on chromosome 20 significantly associated with elevated protein content was identified in the cultivar Danbaekkong. The present research aimed to introgress the Danbaekkong high protein allele into elite lines with different genetic backgrounds by developing and deploying robust DNA markers. A multiparent population consisting of 10 F_5-derived populations with a total of 1115 recombinant inbred lines (RILs) was developed using 'Benning HP' as the donor parent of the Danbaekkong high protein allele. A new functional marker targeting the 321 bp insertion in the gene *Glyma.20g085100* was developed and used to track the Danbaekkong high protein allele across the different populations and enable assessment of its effect and stability. Across all populations, the high protein allele consistently increased the content, with an increase of 3.3% in seed protein. One hundred and three RILs were selected from the multiparent population for yield testing in five environments to assess the impact of the high protein allele on yield and to enable the selection of new breeding lines with high protein and high yield. The results indicated that the high protein allele impacts yield negatively in general, however, it is possible to select high yielding lines with high protein content. An analysis of inheritance of the Chr 20 high protein allele in Danbaekkong indicated that it originated from a *G. soja* line (PI 163453) and is the same as other *G. soja* lines studied. A survey of the distribution of the allele across 79 *G. soja* accessions and 35 *G. max* ancestors of North American soybean cultivars showed that the high protein allele is present in all *G. soja* lines evaluated but not in any of the 35 North America soybean ancestors. These

results demonstrate that *G. soja* accessions are a valuable source of favorable alleles for improvement of protein composition.

Keywords: Soybean, Seed Protein, Danbaekkong, Chromosome 20 QTL, Multiparent populations, Yield

Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most important sources of protein and oil for direct and indirect human use. Soybean oil is omnipresent in the food industry, while soybean meal is the primary source of protein for livestock. Over the past 33 years, soybean yield in the United States increased 40.8%, however, the protein content went in an opposite direction, decreasing from 35.8 to 33.5% (Naeve and Miller-Garvin, 2021). Farmers and grain processors historically have had no incentive to produce and deliver soybeans with high protein and therefore no focus has been given in improving this seed component. The reduction in seed protein has negative effects on soybean value, as lower protein content makes it difficult to meet the requirements of the livestock industry for feed (Brumm and Hurburgh, 1990; de Borja Reis et al., 2020).

The genetic component is a major factor in the determination of seed composition in soybean. Lee et al. (2019) demonstrated the importance of the genetic factors for protein composition (heritability of 0.94) and confirmed the antagonist relationship between protein and oil ($r = -0.75$; $P < 0.0001$). More than 160 protein quantitative trait loci (QTLs) from 35 different studies have been reported (Patil et al., 2017) and one of these QTLs, located on chromosome (Chr) 20, has been repeatedly identified in several studies (Diers et al., 1992; Hwang et al., 2014; Vaughn et al., 2014; Warrington et al., 2015; Qi et al., 2016). This QTL has received the attention of many researchers because of its high additive effect and stability

(Lestari et al., 2013). Warrington et al. (2015) demonstrated that this QTL explained 55 % of the phenotypic variation of seed protein content in a bi-parental population derived from a cross of 'Benning' (PI 595645) (Boerma et al., 1997) and Danbaekkong (PI 619083) (Kim et al., 1996). Danbaekkong is a South Korean cultivar that contributed to high protein content in the population (Warrington et al., 2015).

Despite the negative relationship of protein with oil and yield, there were reports on the feasibility of developing lines with increased protein content and high yield (Cober and Voldeng, 2000; Brzostowski et al., 2017). Prenger et al. (2019) developed Benning HP as a near-isogenic line (NIL) with a high-protein allele on Chr 20 by backcrossing an F_5-derived line from Benning × Danbaekkong to the recurrent parent Benning. This line has high protein content and yield equivalent to the recurrent parent Benning, demonstrating that it is possible to mitigate the negative effects of the high protein allele on yield with progeny selection. However, it is still not clear how the protein and yield relationship work in multiple genetic backgrounds.

A Chr 20 QTL for protein content was detected in the same location of previous mapping studies in a genome-wide association analysis (GWAS) with accessions from the USDA Soybean Germplasm Collection conducted by Vaughn et al. (2014). Bandillo et al. (2015) also analyzed 12,000 accessions from the same collection and identified a protein QTL in the same region. The GWAS hits in these studies were associated with the alleles frequently found in Korean accessions. Using the similar data set, Patil et al. (2017) performed a genome-wide phylogenetic analysis comparing Danbaekkong, North American Soybean Ancestors (NASA), Asian landraces, and several *Glycine soja* lines. When all SoySNP50K SNPs were considered, Danbaekkong was clustered with the NASA, however, when SNPs in the range of 27–32 Mb on Chr 20 were analyzed, Danbaekkong was clustered separately from NASA. This result indicated

that NASA likely have a different allele from Danbaekkong at the Chr 20 and introgression of the high protein allele into elite soybean lines could improve the seed protein content.

Soybean accessions in the USDA Germplasm Collection have great variation for protein content, with accessions reaching up to 57% of seed protein (USDA, 2023). This resource can be tapped to increase the overall protein content and quality in soybean breeding programs. It has been observed that *G. max* cultivars developed in Asia, especially in South Korea, usually have a high content of seed protein than those developed in other countries (Vaughn et al., 2014; Bandillo et al., 2015; Patil et al., 2017). It is likely a result of the historical breeding efforts in that region to focus on the improvement of seed composition for soy food products, such as tofu and soy sauce (Lee et al., 2015). Danbaekkong is a cultivar developed in South Korea based on the selection for seed yield, protein content, quality, and tofu yield (Kim et al., 1996). The Korean accessions with high protein content are an important source of genetic diversity that can be used in U.S. soybean breeding programs to improve nutritional composition.

Recently, a gene was identified underlying control of the protein QTL on Chr 20. Fliege et al. (2022) performed fine mapping in multiple populations using an *G. soja* line (PI 468916) as the QTL donor and narrowed the QTL interval to a region of 77.8 kb. In this interval, a 321 bp fragment was present in the 4th exon of the gene G*lyma.20g085100* in low protein lines. Using a RNAi experiment, the authors demonstrated that the variation in *Glyma.20g085100* was responsible for the difference in protein content. Similarly, Goettel et al. (2022) indicated that *Glyma.20g085100* is the gene responsible for elevated protein at the Chr 20 QTL and soybean lines without the 321 bp insertion exhibit increased protein content, while the lines with the 321 bp insertion had low protein. The authors concluded that the insertion was likely caused by a

transposable element and during the domestication process, the insertion allele is fixed in most *G. max* lines.

In the present research, we aimed to validate the Chr 20 QTL from Danbaekkong for increased protein content; introgress the allele into a wide range of genetic backgrounds for protein improvement; and elucidate the inheritance of the Danbaekkong high protein allele.

Material and Methods

Plant Materials and Population Development

The population consisting of 140 RILs derived from Benning × Danbaekkong originally used to map the Chr 20 QTL was analyzed to saturate the QTL region. The seeds, original phenotypic data and genotypic data were obtained from Warrington et al. (2015). To enable identification of polymorphisms in the QTL region, seven soybean lines with high and low protein content were selected for genome sequencing (Table 2.S1). The elite parent Benning and the high protein parent Danbaekkong (PI 619083) were sequenced together with one high protein *G. soja* accession (PI 163453) and three high protein *G. max* lines (PI 398589, PI 408012, PI 602447) that have a haplotype in the QTL region similar to Danbaekkong. The sequence of the *G. soja* accession PI 468916 that was used in the original mapping study of the Chr 20 QTL by Diers et al. (1992) was obtained from Zhou et al. (2015) and Bayer et al. (2021).

A set of 10 populations was developed by crossing Benning HP with 10 elite lines in 2016 (Table 2.S2). Benning HP is a MG VII near-isogenic line of Benning (PI 595645), carrying the introgression of the Chr 20 high protein allele from Danbaekkong (PI 619083) (Prenger et al., 2019). The populations have a structure of a nested association mapping population, where Benning HP is the hub parent (Supplementary Figure S1).

The F_1 generation was grown in the University of Georgia (UGA) greenhouse in Athens, GA during the winter of 2016-2017. During the summer of 2017, the F_2 generation was grown at the UGA Iron Horse Farm in Watkinsville, GA and then two cycles of single seed descent advancement were conducted to advance the F_3 and F_4 generations during the winter of 2017-2018 in the Puerto Rican nursery. In 2018, the F_5 generation was grown at the UGA Iron Horse Farm and plants from each population were harvested and threshed individually. In the summer of 2019, plant rows were grown in an unreplicated augmented design along with the parents and three commercial check cultivars AG5534, AG6534, and AG7934.

Whole Genome Re-sequencing

The lines selected for sequencing were grown in a greenhouse and leaf tissue was collected three weeks after planting. For each genotype, a bulked sample of 12 plants were collected and leaf tissue was lyophilized and ground. Genomic DNA was extracted using GeneJet Plant Genomic DNA purification mini kit (Thermo Scientific, Boston, MA, USA) and 150 bp DNA fragments were sequenced with NextSeq Sequencing instrument (Illumina, San Diego, CA). Adapters were removed from the raw Fastq files using Trimmomatic v0.36 (Bolger et al., 2014), and sequencing reads were mapped to the soybean genome Wm82.a2.v1 (https://data.jgi.doe.gov) with Bowtie2 v2.3.3.1) (Langmead and Salzberg, 2012). SNP and indel calls were performed with the GATK HaplotypeCaller software (McKenna et al., 2010) and variants were annotated with SNPeff version 4.3t (Cingolani et al., 2012). Variant visualization in the Chr 20 QTL region was performed with the Integrative Genomics Viewer (IGV - v2.9.5) (Robinson et al., 2011).

Marker Design and Genotyping

The RILs from the Benning × Danbaekkong population were planted in the greenhouse and DNA extraction was performed on leaf tissue using the CTAB method (Keim et al., 1988). For the multiparent population, DNA was extracted from seed samples from all 1115 RILs in the 10 populations with a modified Edwards extraction (Edwards et al., 1991). KASP (LGC, Hoddesdon, UK) and TaqMan assays (Applied Biosystems, Foster City, CA) were designed using Geneious Primer version 2021.2 based on polymorphisms present in the QTL region identified from the SoySNP50K data (Song et al., 2013) and whole genome sequence of the seven sequenced soybean lines (Danbaekkong, Benning, PI 163453, PI 398589, PI 408012, PI 602447, and PI 468916) (Tables 2.S3, 2.S4, and 2.S5). The gene specific marker GSM1252 targeting the 321 bp insertion at the gene *Glyma.20g085100* was designed based on information previously published by Fliege et al. (2022) and Goettel et al. (2022).

KASP reactions were performed in a 4μL volume with 2 μL of master mix (1.97 μL KASP 2X and 0.053 μL of primers) and 2 μL of 10–20 ng/μL genomic DNA. Similarly, TaqMan reactions were also conducted in a 4μL volume including 2 μL of master mix (2 μL of TaqMan Universal Master Mix II and 0.2 μL of 5X Custom TaqMan SNP Genotyping Assay) and 2 μL of 10–20 ng/μL genomic DNA. PCR was performed in the BioRad C1000 Touch Thermal Cycler and PCR plates were read in either LightCycler® 480 (Roche, Germany) or TECAN infinite M200 microplate reader (Tecan US, Inc, Durham, NC). Cycling conditions for the KASP assays were 15 min at 94°C, 10 cycles of 15 sec at 94°C and 1 min at 65°C and 30 cycles of 20 sec at 94°C and 1 min at 57°C. Cycling conditions for the TaqMan followed a modified touchdown PCR with an initial 10 min at 95°C, 10 cycles of 20 sec at 95°C and 1 min at 71°C, decreasing 0.5°C each cycle, and 30 cycles of 15 sec at 92°C and 1 min at 58°C.

Diversity Panel

To analyze the distribution of the Chr 20 high protein QTL, 35 North America soybean ancestors (Gizlice et al., 1994) and 79 diverse *G. soja* accessions (La et al., 2019) were genotyped using the gene specific marker GSM1252. The 35 *G. max* soybean ancestors contributed 95% of the genes found in modern soybean cultivars (Gizlice et al., 1994) and the 79 *G. soja* lines are a core set that represent the genetic diversity within the entire USDA *G. soja* Collection (La et al., 2019). These accessions were planted in the greenhouse and leaf tissue was collected two weeks after planting. DNA extraction was performed with the CTAB method (Keim et al., 1988).

The seed composition of the 79 *G. soja* accessions was obtained from La et al. (2019) and the data for 25 of 35 North American Soybean ancestors was collected with Near-Infrared Spectroscopy Perten DA 7250 Analyzer (PerkinElmer Inc., Waltham, MA, USA) from seeds harvested in the USDA winter nursery in Puerto Rico in 2018. The phenotypes of the remaining 10 accessions were retrieved from USDA GRIN (https://npgsweb.ars-grin.gov/gringlobal/).

Another panel of 35 *G. soja* lines was assembled to compare the genome sequence variation at the gene level and survey the distribution of the high protein allele. The raw sequencing data was generated in previous studies (Bayer et al., 2021; Valliyodan et al., 2021) and available at the Short Read Archive (SRA) database at NCBI (www.ncbi.nlm.nih.gov). Adapters were removed from the raw Fastq files using Trimmomatic v0.36 (Bolger et al., 2014) and sequencing reads were mapped to the soybean genome Wm82.a2.v1 with Bowtie2 v2.3.3.1 (Langmead and Salzberg, 2012).

Yield Trials of Selected RILs

To understand the effects of the Danbaekkong high protein allele on yield in different genetic backgrounds, a set of RILs from the multi-parent populations with high and normal protein content were selected for evaluation in yield trials. A total of 103 lines were planted in three locations in 2020 and 2021. In 2020, all the lines were grown in a randomized complete block design with two replications per location and each line was planted in a 2-row plot with a length of 4.9-m spaced by 76.2-cm and a planting density of 27 seeds meter^{-1}. Forty-six lines were selected based on yield and agronomic performance and grown in 2021 in a randomized complete block design with three replications in a 4-row plot with the same plot length and row spacing. Agronomic practices followed the recommended guidance for soybean production in Georgia (Bryant, 2020). All plots were end-trimmed before harvest to avoid edge effect, resulting in a length of 3.7 m. The two center rows were harvested, and weight and moisture were measured on combines. Approximately 200 seeds were sampled from each plot for seed composition analysis.

Seed Composition Analysis

The contents of protein and oil were determined using the NIR Perten DA 7250 Analyzer (PerkinElmer Inc., Waltham, MA, USA) and the instrument was calibrated by the manufacturer using thousands of samples with known seed composition values for whole seed and ground seed samples. Seed composition was reported on a dry matter basis. Analysis of the multiparent population was performed on the seeds from single plants in 2018 and from the plant rows in 2019. For the yield trials in 2020 and 2021, samples of 200 seeds were obtained from each plot.

Statistical and QTL analyses

Phenotypic and genotypic data was analyzed in RStudio (R version 3.4.4) using the packages lme4 (Bates et al., 2015) and BreedR (Muñoz and Rodriguez, 2020) and data visualization was created with ggplot2 (Wickham, 2016). The phenotypic values for the QTL analysis of the multi-parent population were calculated by fitting a model with the subpopulation and year effects as fixed and the genotype effect as random. For the data from the Benning × Danbaekkong RIL population, best linear unbiased predictions (BLUPs) were obtained by fitting a model with the environment (location + year) as fixed and genotype effect and replication as random effect. Analysis of the phenotypic data from yield trials with the 103 selected breeding lines was performed by fitting a model with the QTL within each subpopulation as a fixed effect, and genotype, environment, and replication as random effects.

To saturate the QTL region identified in the Benning × Danbaekkong RIL population, additional markers were developed in the QTL interval based on comparison of the sequencing data from the seven sequenced genotypes. Linkage map construction and QTL analysis were performed with the R package R/qtl (Broman et al., 2003). Associations between markers and protein content were established with a regression function using a LOD significance threshold determined by 1000 permutations. Recombination distances were calculated using Kosambi's mapping function with simple interval and composite interval mapping methods in the QTL position estimation. To understand the effects of the QTL in a broad genetic background, a multi-parent population QTL analysis was performed using an R package mppR (Garin et al., 2020). In each round of mapping, the population was randomly partitioned into five subsets and one of the subsets was used for validation of the parameters calculated in the other four subsets. Composite interval mapping was performed in each subset 100 times and the QTL position was determined by the location of the most significant marker across all iterations.

Results

Danbaekkong high protein allele

The RIL population derived from the Benning × Danbaekkong cross (N=140) was genotyped with the Chr 20 QTL flanking markers previously used by Warrington et al. (2015) and 17 new additional markers designed based on variants found in the comparison of the seven sequenced lines. The markers were combined to saturate the Chr 20 QTL region and one of the markers used (GSM1252) specifically targeted the 321 bp insertion in the gene *Glyma.20g085100* identified by Fliege et al. (2022). The QTL interval was identified in a genomic region between 27.7 and 33.0 Mb across all environments tested and the marker GSM1252 designed from the gene *Glyma.20g085100* was one of the most significant markers across the environments (Figure 1). In this population, the QTL explained 47.5% of the phenotypic variation and had an additive effect of 1.3% in the protein content. The homozygous RILs for the low protein allele at the GSM1252 locus had an average protein content of 43.8 %, while the lines with homozygous high protein allele had the protein content of 46.4 %.

QTL mapping was also performed in the multi-parent population and the QTL region was identified to the interval between 31.8 and 32.2 Mb (markers GSM1252 and GSM0455). This region is located within the QTL interval identified in the analysis of the Benning × Danbaekkong RIL population (Figure 2.2). After the estimation of the QTL parameters, an association analysis between the *Glyma.20g85100* marker (GSM1252) and the content of protein and oil was performed using the lines in the multiparent population. The high protein allele from Danbaekkong was associated with an increase in the protein of 3.3% on average (ranging from 2.6 to 3.7%) across all populations. The highest increase in protein content was observed in population G13-6299 × Benning HP, with protein content going from 40.8 to 44.5%. The highest

average value of protein obtained was in the population Woodruff × Benning HP with lines carrying the high protein allele reaching 45.4% (Figure 2.3, Figure 2.4, and Table 2.S6).

The increase in the protein content was accompanied by a reduction in oil content in all populations, ranging from a reduction of 1.4% in Benning HP × G10PR-56444R2 to 2.0% in N10-711 × Benning HP and Benning HP × G11PR-56238R2. On average, for every 1.8% increase in protein, there was a decrease of 1% in oil. The populations N08-174 × Benning HP and Benning HP × G10PR-56444R2 had an average oil content ≥ 20% and protein content ≥43.5%, demonstrating the possibility of having high protein and oil above 20%. (Table 2.S6).

The fact that Benning HP was used either as a male or female parent in the multiparent population, enabled evaluation of any maternal effect of the Danbaekkong high protein allele. It was observed that the Danbaekkong high protein allele increased the protein in a similar magnitude either having the Benning HP as the female (44.5%) or the male (44.3%) parent in the cross (Table 2.S6).

Effects of the Danbaekkong high protein allele on yield

To assess the effects of the protein QTL on yield, 103 RILs were selected from the multiparent population based on agronomic performance and visual assessment of plant appearance, lodging, and maturity to enter the 2020 and 2021 yield trials. Population N10-711 × Benning HP had the highest number of lines in the trials (27 in 2020 and 13 in 2021), while Benning HP × G11PR-56238R2 had the lowest number, with three lines in the yield trials. Overall, all pedigrees had lines with the high protein allele or low protein allele variant evaluated in both years, except for Benning HP × G11PR-56238R2 population, which was evaluated only in 2020 and Benning HP × G10PR-56444R2 did not have lines with the high protein allele tested. Having lines with and without the Danbaekkong high protein allele evaluated in yield

trials in 9 of the 10 pedigrees enabled a comparison of the effects of the increased protein content on yield in multiple genetic backgrounds.

In the yield trials, the lines carrying the high protein allele had a consistently higher protein content across all the populations, with an average increase of 2.0% in protein content. The only exception was population R12-514 × Benning HP, in which lines with the high protein allele in population did not have a significant increase in the protein content. Population G13-6299 × Benning HP had the highest increase in protein, from 40.1 to 43.2% and population N10-711 × Benning HP had the highest average value of protein, with 43.8% (Figure 2.5a). The oil content had an overall reduction of 1%, but variation was observed across the different populations, ranging from a 2% reduction in Benning HP × G11PR-56238R2 to no detectable reduction in R12-514 × Benning HP (Figure 2.5b). In the comparison of the protein production per hectare, the populations also had different performance. Lines with the high protein allele from the population N10-711 × Benning HP had an increase of 94 kg ha^{-1} in protein production, but in the population N05-7432 × Benning HP the lines with the high protein allele had a decrease of 218 kg ha^{-1}. When considering the performance of all populations together, there was no difference (p=0.41) in the protein production per hectare in the lines with or without the Danbaekkong high protein allele, 2048 vs 2080 kg ha^{-1}, respectively (Figure 2.5c, Table2.S7).

Overall, the high protein negatively impacts the yields, with an average reduction of 313 kg ha^{-1}. However, there was variation across the different populations, with population N05-7432 × Benning HP having a yield reduction of 719 kg ha^{-1} to the population N10-711 × Benning HP with a yield reduction of only 55 kg ha^{-1} in the lines with the high protein allele. Of the 103 lines evaluated, 20 lines from different populations had yield similar or higher than commercial check AGS 738RR and 14 of these lines had protein content higher than 40% (Table

2.S8). The line G19-11395 from population N05-7432 × Benning HP did not have the high protein allele but stood out with the highest overall yield, 5880 kg ha^{-1}, 13.8% higher but not significantly different from AGS-738RR. The line G19-11191 from population Woodruff × Benning HP was the only line carrying the high protein allele (43.6% protein) that had yield comparable to the AGS 738RR (100.4%), with 5189 kg ha^{-1}. Other three lines, G19-11422 (N05-7432 × Benning HP), G19-11111 (G13-6299 × Benning HP) and G19-2139R2 (Benning HP × G11PR-56151R2) carrying the high protein allele at GSM1252, had protein content exceeding 43% and yielded >95% of AGS 738RR (Table 2.S8). These results exemplify the possibility of combining high yield with improved seed composition.

Effect of maturity on seed protein

The association between maturity and the high protein alleles was evaluated across the different pedigrees in the multiparent population. Nine out of 10 populations studied had the lines carrying the high protein allele reaching maturity earlier than those with normal protein. Overall, high protein lines reached maturity 3.5 days earlier than those with low protein (Table 1.1). The population with the biggest difference was N05-7432 × Benning HP, in which the lines having the high protein allele matured 6.1 days earlier than those with the low protein allele. On the other hand, there was no significant difference in maturity between lines with the high protein allele and those with the low protein allele in the population N08-174 × Benning HP.

Distribution of the Chr 20 high protein allele among the soybean ancestors and G. soja lines

The presence of the Danbaekkong high protein allele was surveyed using the gene marker GSM1252 in a panel of 35 *G. max* ancestral lines that contributed 95% of the genes found in modern soybean cultivars (Gizlice et al., 1994). These lines provided a good opportunity to understand the distribution of the high protein allele in the North America soybean breeding

pool. The results indicated that all 35 *G. max* ancestors have the low protein allele at the gene *Glyma.20g085100*, and the average protein content was 41.6% (ranging from 38.1 to 45.7 %) (Table 2.2, Figure 2.6).

Another analysis was performed to study the distribution of the high protein allele across *G. soja* accessions. A panel of 79 diverse *G. soja* that represent the genetic diversity in USDA Soybean Germplasm Collection was surveyed (La et al., 2019). All the *G. soja* lines evaluated presented the high protein allele on the Chr 20 and had an average protein content of 44.4% (ranging from 39.8 to 49.4%) (Table 2.3, Figure 2.6). To confirm the presence of the high protein allele in *G. soja*, the sequence of 35 accessions that have not been studied previously was analyzed for the presence of the insertion in *Glyma.20g085100*. Confirming the previous results, all *G. soja* lines evaluated have the high protein allelic variant (Table 2.S9).

Discussion

Danbaekkong high protein allele

Using new molecular markers positioned in the interval where the protein QTL has been repeatedly identified (29.8 to 34.3 Mb), genotyping was performed in the Benning × Danbaekkong RIL population (N=140) and in a multiparent population (N=1115). Of these markers, GSM1252 was developed based on previous research that identified *Glyma.20g085100* controlling the protein at the Chr 20 QTL (Fliege et al., 2022; Goettel et al., 2022). GSM1252 was developed as TaqMan marker with one probe targeting the flanking regions of the insertion aiming to capture the alleles without the 321 bp insertion and another probe that binds to a fragment of the insertion and the right flanking site (Figure 2.S2). Overall, the marker exhibited a good performance in separating the lines with and without the insertion and it is a useful tool to select lines for high protein. The QTL analysis confirmed the variation in *Glyma.20g085100* to

be associated with protein content in the populations derived from Danbaekkong. However, instead of GSM1252, marker GSM1122 was the most significant marker at the locus. This can be attributed to the fact that Chr 20 QTL is a region of strong linkage disequilibrium (Vaughn et al. 2014).Therefore, if a SNP marker has a slightly better genotyping performance, it will have a better association. The data analysis in Benning × Danbaekkong and the multiparent population indicated a confidence interval of 503,806 bp between the markers GSM1252 and GSM0455 (31,778,817bp – 32,282,623) (Wm82.a2.v1). This region overlaps perfectly with previously published mapping work that identified the Chr 20 QTL (Bolon et al., 2010; van Warrington et al., 2014; Vaughn et al., 2014; Lee et al., 2019; Wang et al., 2021). In the analysis of the multiparent mapping population, the flat QTL peak in the region between 31.8 and 32.8 Mb indicated that this genomic region has a large linkage disequilibrium block.

To elucidate the origins of the Danbaekkong high protein allele, an analysis of the Danbaekkong pedigree was conducted. One of the Danbaekkong's parent is the cultivar Dongsan 69 from South Korea and the pedigree of this cultivar is unknown since no release information is available. The other parent is D76-8070 which is an MG V line developed by Edgar Hartwig in his effort to breed soybean cultivars with increased protein content (Hartwig, 1990). D76-8070 was developed through the selection of progeny from multiple crosses ('Hill' × 'Sioux', FC 31745 × D49-2510, Hill × PI 96983, and D49-24914 × PI 163453). The progeny from each of these crosses were selected for disease resistance, agronomic traits, and high protein content (> 45%) and the selected lines were intercrossed to develop D76-8070 (Figure 2.S3). PI 163453 is the only *G. soja* line present in the pedigree of D76-8070 and was hypothesized as the origin for the high protein QTL. To verify this hypothesis, the haplotypes of PI 163453 and Danbaekkong were compared using the 6,353 SNPs between 30 and 34 Mb on Chr 20 called from the

sequencing data. The genetic similarity analysis showed that the Danbaekkong haplotype at the Chr 20 QTL region is 99.95% identical to PI 163453 (Table 2.S10). To confirm the inheritance of the protein QTL, D76-8070 was also genotyped with GSM1252 and the results indicated that it carries the same allele as PI 163453, Danbaekkong and Benning HP (Table 2.S11).

To quantify the Chr 20 fragment that was transferred from PI 163453 to Danbaekkong, 408 SNPs from the SoySNP50K SNP dataset distributed along the Chr 20 were used and it was observed that the PI 163453 fragment that was transferred to D76-8070 spans from 21 to 34.6 Mb and the D76-8070 fragment that was transferred to Danbaekkong, starts at 2 Mb, and ends at 36 Mb. Subsequently, a fragment from 0.2 to 37 Mb from Danbaekkong was transferred to Benning HP (Figure 2.S4). These results indicate that the high protein allele is originally from PI 163453, and it was transferred to D76-8070 through the work of Hartwig. Then D76-8070 was used in South Korea to develop Danbaekkong, which eventually returned to the United States and was used to develop the isogenic line Benning HP.

The haplotype of PI 163453 was also compared to the *G. soja* line PI 468916 used in the mapping study that identified the Chr 20 QTL (Diers et al., 1992).The comparison revealed that PI 163453 is only 43% similar to PI 468916 when considering all the SNPs in 30-34Mb window, but when comparing the sequence of the gene *Glyma.20g085100*, it was observed that PI 163463 is also missing the 321 bp fragment as PI 468916 (Table 2.S11; Figure 2.S5). These results indicate that although PI 163453 and PI 468916 are different at the haplotype level, they carry the same high protein allele in *Glyma.20g085100*.

Goettel et al. (2023) indicated that the *Glyma.20g085100* high protein allele was transferred from *G. soja* to *G. max* in three independent events likely during the process of domestication in East Asia. In the present research it was demonstrated that the Danbaekkong

high protein allele came from the intentional introgression conducted by Edgar Hartwig where the *G. soja* PI 163453 was used as a grand parent to develop D76-8070 (Hartwig, 1990). Analyzing the haplotypes in the *Glyma.20g085100* region (Chr20, 29 - 34 Mb) revealed that both PI 163453 and Danbaekkong were grouped into the cluster 3 identified by Goettel et al. (2023)(Figure 2.7). Cluster 3 is predominantly composed of the accessions from Chinaexcept Danbaekkong that is a derived progeny from PI 163453.

Distribution of the Chr 20 high protein allele among the soybean ancestors and G. soja lines

An analysis of the distribution of the high protein allele was performed using 35 *G. max* that represents the diversity of the North America soybean cultivars (Gizlice et al. 1994). The results indicated that none of the 35 *G. max* ancestors carry the high protein allele in *Glyma.20g085100.* However, three soybean ancestors, CNS (PI 548445), Arksoy (PI 548438) and Bilomi No. 3 (PI 240664) have protein content higher than 44% but do not carry the Chr 20 high protein allele. CNS, Arksoy and Bilomi No. 3 were originally collected in China, North Korea, and Philippines, respectively, and it is possible that these three accessions harbor protein QTLs in other genomic regions. To our knowledge these ancestors have not been used in QTL mapping studies yet and they could reveal more information about the genetic control of protein in soybean.

Soybean lines with protein content reaching values of 47.2% have been developed (Wilcox and Cavins, 1995) and some lines have been released as cultivars in the United States in an effort to improve the seed composition, such as Protana, with 43% protein (Probst et al., 1971), Prolina, with 46% protein (Burton et al., 1999) and Prohio with 44.1% (Mian et al., 2008). More recently, soybean breeders focused on combining high yield and improved protein content and several breeding lines have been released. Chen et al. (2017) developed UA 5814HP as a

new soybean cultivar with high seed protein content (45.5%) and yield comparable to elite checks. Pantalone and Smallwood (2018) released TN11-5102 as a high yield and high protein line with 42% protein. Shannon et al. (2022) developed S09-13185, with 44% protein content and Li et al. (2022) released G11-7013 with a protein content of 43.6%. Despite these efforts, the proportion of high protein lines in North American germplasm is low. According to Patil et al. (2017), most soybean cultivars in the United States are fixed for the low protein allele at the Chr 20 locus, and the introgression of the high protein allele have the potential to improve the seed protein content in soybean cultivars in North America.

Goettel et al. (2022) analyzed a panel of 398 *G. max* (259 Cultivars and 139 Landraces) and 150 *G. soja* accessions from the USDA Soybean Germplasm Collection and observed that only 21 *G. max* lines had the Chr 20 high protein allele. Of these 21 *G. max* lines that have the high protein allele, one line was from India, two from Japan, four from China and 14 lines from South Korea, where Danbaekkong originated. Eight of the 14 Korean lines are cultivars with yellow seed coat, indicating that the Chr 20 high protein allele has been selected and used in the development of soybean cultivars in Korean breeding programs. Lee et al. (2015) conducted a pedigree reconstruction of Korean soybean varieties and demonstrated that since 1913, soybean breeding programs have focused primarily on the improvement of seed protein composition for processing as soy food, such as soy sauce and tofu.

Differently from *G. max*, it was observed that all 79 *G. soja* from the USDA core collection analyzed carry the high protein allele at *Glyma.20g085100*. When analyzing the sequence of additional 35 *G. soja* accessions, all of them also carry the high protein allele. In a similar way, Goettel et al. (2022) analyzed a panel of 150 *G. soja* accessions and found that 147 lines had the high protein allele confirmed. Due to the widespread presence in *G. soja* of the high

protein allele in *Glyma.20g085100* and the low frequency in *G. max*, and the fact that *G. soja* is the closest ancestor to *G. max*, it is possible to infer that the high protein allele is the original state of the gene.

Effects of the Danbaekkong high protein allele

A single marker analysis with the *Glyma.20g085100* marker was performed to understand the stability and effect of the gene across different genetic backgrounds. The analysis revealed that the high protein allele inherited from Danbaekkong increased the protein by 3.3% on average (ranging from 2.6 to 3.7%) across all 10 populations tested in 2018 and 2019. The increase in protein content was also observed in the yield trials conducted in 2020 and 2021. In these trials the high protein allele had an average increase of 2.0% in the protein and only population R12-514 × Benning HP did not show a significant increase in protein. This protein increase is similar to the estimate by Brzostowski et al. (2017), when the introgression of the Danbaekkong allele into two soybean lines, caused an increase of 2 % across four environments. The present results are close to the estimates by Warrington et al. (2015), where the author indicated a gain of 2.7 % in protein with the Danbaekkong allele.

One of the well-known effects of the increase of protein content is the reduction of oil (Cober and Voldeng, 2000; Chung et al., 2003; Vaughn et al., 2014; Patil et al., 2018). According to Hanson et al. (1961), this relationship is dictated by a ratio of 2:1, in which the energy demanded to synthesize 2 protein units corresponds to 1 unit of oil. Other studies have shown that the protein to oil ratio is between 1.5 to 1.7 (Hartwig and Kilen, 1991; Chung et al., 2003). In the present research, it was observed that for every 1% increase in protein, there was a decrease of 0.55% in oil, representing a ratio of 1.8:1.

Several studies have indicated a negative relationship between protein and yield, with correlation values reaching up to -0.62 (Cober and Voldeng, 2000; Sebolt et al., 2000; Cunicelli et al., 2019). Overall, a negative correlation between these two traits appears to be common, but contrary to the omnipresent antagonist relationship between protein and oil, protein and yield do not have a consistent correlation when comparing multiple environments (Wilcox and Cavins, 1995; Prenger et al., 2019). In the present research, lines with the high protein allele in general yield 313 kg ha^{-1} less (55 to 719 kg ha^{-1}) than those with the low protein allele within the same population. Brzostowski et al. (2017) found a yield reduction ranging from −273 to −558 kg ha^{-1} when introgressing the Danbaekkong allele into two soybean lines. In the same way, Goettel et al. (2022) indicated that the low protein allele at *Glyma.20g085100* is associated with a yield increase of 150.3 kg ha^{-1}. Despite the negative effect of the high protein allele on yield, it was possible to identify lines carrying the high protein allele (>43% protein) with comparable yield to the commercial checks (>95% yield). This shows that there is potential to couple high yield and high protein content with selection during breeding, and the negative association between protein and yield can be minimized.

An association between the presence of the high protein allele in *Glyma.20g085100* and maturity was observed across different populations, where lines with the high protein allele matured approximately 3.7 days earlier than their counterparts in the same population. Similar results were found by Prenger et al. (2019), where lines carrying the Danbaekkong allele matured earlier than those without the allele. The gene *Glyma.20g085100* is located 1.4 Mb upstream of the maturity locus E4 (Liu et al., 2008). Since Danbaekkong is an MG V cultivar, it is possible that it possesses the early maturity allele at the *E4* locus linked with the high protein allele in a

coupling phase. Therefore, the difference in maturity in lines with high protein derived from Danbaekkong is due to linkage between the high protein QTL and the maturity gene *E4*.

To our knowledge, the present study was the first time a QTL for protein content in soybean has been fully assessed in a wide variety of genetic background simultaneously with several environments of yield trials and its breeding history from *G. soja* to *G. max* described. This study complements and validates the findings of previous research about the role of *Glyma.20g085100* in determining the protein content in soybeans, providing more information about the effects and stability of the QTL, and confirming the value of its use to improve soybean seed composition.

Conclusions

In this research, a gene specific marker was designed for *Glyma.20g085100 (GSM1252)* and genotyping the bi-parental and multi-parental populations confirmed the effectiveness of this marker as well as other flanking markers. This information can be useful resources for breeding programs to introgress the high protein allele into elite lines. The analysis of the distribution of the *Glyma.20g085100* alleles revealed that the 35 *G. max* accessions that represent the genetic diversity of North American soybean cultivars have the low protein allele, while the 79 *G. soja* accessions surveyed possess the high protein allele. The analysis of the pedigree of Danbaekkong indicated that its high protein allele was inherited from *G. soja* PI 163453, which is the same as the one from PI 468916. The Danbaekkong high protein allele increased the protein content in all populations tested in 2018 and 2019 with average of 3.3%, ranging from 2.6 % in Benning HP × G10PR-56444R2 to 3.7 % increase in G13-6299 × Benning HP. In the yield trials in 2020 and 2021, the allele increased the protein in 2% on average and was stable across multiple environments. It was observed that the increase in protein was accompanied by an overall

decrease in oil and yield. However, it was possible to select breeding lines with the high protein allele and yield comparable to elite checks and this will enable the development of new cultivars with high protein content and high yield.

Acknowledgments

We thank Tatyana Nienow, Nicole Bachleda, Dale Wood, Brice Wilson, and Brian Little at the University of Georgia for the technical support.

References

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

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1–48. doi: 10.18637/jss.v067.i01.

Bayer, P. E., Yuan, Y., Batley, J., Nguyen, H. T., Valliyodan, B., Varshney, R. K., et al. (2021). Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome*, 1–12. doi: 10.1002/tpg2.20109.

Boerma, H. R., Hussey, R. S., Phillips, D. V., Wood, E. D., Rowan, G. B., and Finnerty, S. L. (1997). Registration of 'Benning' Soybean. *Crop Science* 37, 1982–1982. doi: 10.2135/cropsci1997.0011183x003700060061x.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bolon, Y. T., Joseph, B., Cannon, S. B., Graham, M. A., Diers, B. W., Farmer, A. D., et al. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biology* 10. doi: 10.1186/1471-2229-10-

41.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112.

Brumm, T. J., and Hurburgh, C. R. (1990). Estimating the processed value of soybeans. *Journal of the American Oil Chemists` Society*. 67, 302–307. doi: 10.1007/BF02539680.

Bryant, C. (2020). *Soybean Production in Georgia*. 1st ed. University of Georgia Cooperative Extension: Athens.

Brzostowski, L. F., Pruski, T. I., Specht, J. E., and Diers, B. W. (2017). Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. *Theorerical and Applied Genetetics* 130, 2315–2326. doi: 10.1007/s00122-017-2961-x.

Burton, J. W., Carter, T. E., and Wilson, R. F. (1999). Registration of ` Prolina ' Soybean. *Crop Science* 39, 1993–1994. doi: 10.2135/cropsci1999.0011183X003900010066x.

Chen, P., Florez-Palacios, L., Orazaly, M., Manjarrez-Sandoval, P., Wu, C., Rupe, J. C., et al. (2017). Registration of 'UA 5814HP' Soybean with High Yield and High Seed-Protein Content. *Journal of Plant Registrations* 11, 116–120. doi: 10.3198/jpr2016.09.0046crc.

Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B., et al. (2003). The Seed Protein, Oil, and Yield QTL on Soybean Linkage Group I. *Crop Science* 43, 1053–1067. doi: 10.2135/cropsci2003.1053.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3. *Fly.* 6, 80–92. doi: 10.4161/cam.20753.

Cober, E. R., and Voldeng, H. D. (2000). Developing high-protein, high-yield soybean populations and lines. *Crop Science* 40, 39–42. doi: 10.2135/cropsci2000.40139x.

Cunicelli, M. J., Bhandari, H. S., Chen, P., Sams, C. E., Mian, M. A. R., Mozzoni, L. A., et al. (2019). Effect of a Mutant Danbaekkong Allele on Soybean Seed Yield, Protein, and Oil Concentration. *Journal of the American Oil Chemists' Society* 96, 927–935. doi: 10.1002/aocs.12261.

de Borja Reis, A. F., Tamagno, S., Moro Rosso, L. H., Ortez, O. A., Naeve, S., and Ciampitti, I. A. (2020). Historical trend on seed amino acid concentration does not follow protein changes in soybeans. *Scientific Reports* 10, 1–10. doi: 10.1038/s41598-020-74734-1.

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

Edwards, K., Johnstone, C., and Thompson, C. (1991). A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Research* 19, 1349. doi: 10.1093/nar/19.6.1349.

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 quantitative trait loci on soybean chromosome 20. *Plant Journal* 1–15. doi: 10.1111/tpj.15658.

Garin, V., Malosetti, M., and van Eeuwijk, F. (2020). Multi-parent multi-environment QTL analysis: an illustration with the EU-NAM Flint population. *Theoretical and Applied Genetics* 133, 2627–2638. doi: 10.1007/s00122-020-03621-0.

Gizlice, Z., Carter, T. E., and Burton, J. W. (1994). Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sciences* 34, 1143–1151. doi:

10.2135/cropsci1994.0011183X003400050001x.

Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., et al. (2022). POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nature Communications* 13, 3051. doi: 10.1038/s41467-022-30314-7.

Hanson, W. D., Leffel, R. C., and Howell, R. W. (1961). Genetic analysis of energy production in the Soybean. *Crop Science* 1, 121–126. doi: 10.2135/cropsci1961.0011183X000100020011x.

Hartwig, E. E. (1990). Registration of soybean high-protein germplasm line 'D76-8070.' *Crop Science* 30, 764–765. doi: 10.2135/cropsci1990.0011183x003000030092x.

Hartwig, E. E., and Kilen, T. C. (1991). Yield and composition of soybean seed from parents with different protein, similar yield. *Crop Science* 31, 290–292. doi: 10.2135/cropsci1991.0011183x003100020011x.

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, 1–12. doi: 10.1186/1471-2164-15-1.

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

Kim, S.-D., Hong, E.-H., Kim, Y.-H., Lee, S.-H., Seong, Y.-K., Park, K.-Y., et al. (1996). A new high protein and good seed quality soybean variety "Danbaegkong". *RDA Journal of Agricultural Sciences* 38, 228–232.

La, T., Large, E., Taliercio, E., Song, Q., Gillman, J. D., Xu, D., et al. (2019). Characterization of select wild soybean accessions in the USDA germplasm collection for seed composition and agronomic traits. *Crop Science* 59, 233–251. doi: 10.2135/cropsci2017.08.0514.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9, 357–359. doi: 10.1038/nmeth.1923.

Lee, C., Choi, M., Kim, H., Yun, H., Lee, B., Chung, Y., et al. (2015). Soybean [*Glycine max* (L.) Merrill]: Importance as a crop and pedigree reconstruction of Korean varieties. *Plant Breeding and Biotechnology* 3, 179–196. doi: 10.1016/S0828-282X(08)70684-6.

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. *Theoretical and Applied Genetics* 132, 1639–1659. doi: 10.1007/s00122-019-03304-5.

Lestari, P., Van, K., Lee, J., Kang, Y. J., and Lee, S.-H. (2013). Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. *Frontiers in Plant Sciences* 4, 1–8. doi: 10.3389/fpls.2013.00176.

Li, Z., Bachleda, N., Wilson, B., Wood, E. D., Buck, J. W., Carter, T. E., et al. (2022). Registration of G11-7013 soybean germplasm with high meal protein and resistance to soybean cyst nematode, southern root-knot nematode, and stem canker. *Journal of Plant Registrations* 16, 430–437. doi: 10.1002/plr2.20204.

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., and Abe, J. (2008). Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics* 180, 995–1007. doi: 10.1534/genetics.108.092742.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research.* 20, 1297–1303. doi: 10.1101/gr.107524.110.20.

Mian, M. A. R., Cooper, R. L., and Dorrance, A. E. (2008). Registration of ‘Prohio’ Soybean.

Journal of Plant Registrations 2, 208–210. doi: 10.3198/jpr2007.09.0531crc.

Muñoz, F., and Rodriguez, L. S. (2020). BreedR: Statistical methods for forest genetic resources analysts. Available at: https://github.com/famuvie/breedR.

Naeve, S., and Miller-Garvin, J. (2021). United States soybean quality - Annual Report. Dep. of Agronomy, University of Minnesota, St. Paul.

Pantalone, V., and Smallwood, C. (2018). Registration of 'TN11-5102' Soybean cultivar with high yield and high protein meal. *Journal of Plant Registrations* 12, 304–308. doi: https://doi.org/10.3198/jpr2017.10.0074crc.

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. *Theoretical and Applied Genetics* 130, 1975–1991. doi: 10.1007/s00122-017-2955-8.

Patil, G., Vuong, T. D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C., et al. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. *Plant Biotechnology Journal* 16, 1939–1953. doi: 10.1111/pbi.12929.

Prenger, E. M., Yates, J., Mian, M. A. R., Buckley, B., Boerma, H. R., and Li, Z. (2019). Introgression of a high protein allele into an elite soybean cultivar results in a high-protein near-isogenic line with yield parity. *Crop Science* 59, 2498–2508. doi: 10.2135/cropsci2018.12.0767.

Probst, A. H., Laviolette, F. A., Athow, K. L., and Wilcox, J. R. (1971). Registration of Protana Soybean. *Crop Science* 11, 312–312. doi: 10.2135/cropsci1971.0011183x001100020050x.

Qi, Z., Pan, J., Han, X., Qi, H., Xin, D., Li, W., et al. (2016). Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments

based on a high-density map. *Molecular Breeding* 36, 1–16. doi: 10.1007/s11032-016-0475-x.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nature Biotechnology* 29, 24–26. doi: 10.1038/nbt.1754.

Sebolt, A. M., Shoemaker, R. C., and Diers, B. W. (2000). Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Science* 40, 1438–1444. doi: 10.2135/cropsci2000.4051438x.

Shannon, G., Chen, P., Crisel, M., Smothers, S., Clubb, M., Vieira, C. C., et al. (2022). S09-13185: High-yield soybean germplasm with elevated protein concentration. *Journal of Plant Registrations* 16, 417–422. doi: 10.1002/plr2.20169.

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, 1–12. doi: 10.1371/journal.pone.0054985.

USDA (2023). Germplasm Resources Information Network (GRIN) - National Plant Germplasm System. Available at: https://www.ars-grin.gov/. Accessed on March 7, 2023

Valliyodan, B., Brown, A. V., Wang, J., Patil, G., Liu, Y., Otyama, P. I., et al. (2021). Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. *Scientific Data* 8, 1–9. doi: 10.1038/s41597-021-00834-w.

Van, K., Hwang, E. Y., Kim, M. Y., Park, H. J., Lee, S. H., and Cregan, P. B. (2005). Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. *Journal of Heredity* 96, 529–535. doi: 10.1093/jhered/esi069.

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 Genes, Genomes, Genetics* 4, 2283–2294. doi: 10.1534/g3.114.013433.

Wang, J., Mao, L., Zeng, Z., Yu, X., Lian, J., Feng, J., et al. (2021). Genetic mapping high protein content QTL from soybean 'Nanxiadou 25' and candidate gene analysis. *BMC Plant Biology* 21, 1–13. doi: 10.1186/s12870-021-03176-2.

Warrington, C., Abdel-Haleem, H., Orf, J. H., Killam, A. S., Bajjalieh, N., Li, Z., et al. (2014). Resource allocation for selection of seed protein and amino acids in soybean. *Crop Science* 54, 963–970. doi: 10.2135/cropsci2013.12.0799.

Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S., et al. (2015). QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. *Theoretical and Applied Genetics* 128, 839–850. doi: 10.1007/s00122-015-2474-4.

Wickham, H. (2016). *ggplot2. Elegant Graphics for Data Analysis.* Springer-Verlag, New York. doi: 10.1002/wics.147.

Wilcox, J. R., and Cavins, J. F. (1995). Backcrossing high seed protein to a soybean cultivar. *Crop Science* 35, 1036–1041. doi: 10.2135/cropsci1995.0011183X003500040019x.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* 33, 408–414. doi: 10.1038/nbt.3096.

Table 2.1. Effects of the high protein QTL on maturity.

Population	Pedigree	Maturity†		Difference
		HP lines	LP lines	
P1	G13-6299 × Benning HP	52.8	55.3	2.5*
P2	Woodruff × Benning HP	53.0	55.5	2.5***
P3	N10-711 × Benning HP	47.8	52.3	4.5***
P4	N05-7432 × Benning HP	51.1	57.2	6.1***
P5	N11-7046 × Benning HP	48.7	52.0	3.3**
P6	N08-174 × Benning HP	41.4	41.9	0.5^{ns}
P7	R12-514 × Benning HP	40.7	44.2	3.4*
P8	Benning HP × G10PR-56444R2	52.4	55.3	2.9***
P9	Benning HP × G11PR-56151R2	48.7	54.3	5.6***
P10	Benning HP × G11PR-56238R2	48.4	53.5	5.1***

Note: HP indicates the high protein allele and LP indicates the low protein allele at GSM1252
*, **, *** indicated significant differences at the 0.05, 0.01 and 0.001 probability level, respectively.
† Maturity is indicated as days after August 31.

Table 2.2. Distribution of the low protein allele (321bp Insertion) among North American soybean ancestral lines as defined by Gizlice et al. (1994).

ID	Origin	MG	Protein (%)	Oil (%)	GSM1252†
FC 31745	-	VI	40.2	21.5	LP
FC 33243	-	IV	38.1	22.5	LP
PI 180501‡	Germany	0	39.1	21.3	LP
PI 240664‡	Philippines	X	44.8	21.1	LP
PI 360955B‡	Sweden	0	42.7	18.2	LP
PI 438471	Sweden	0	38.2	20.3	LP
PI 438477	Sweden	0	39.6	19.7	LP
PI 548298	China	III	43.0	19.9	LP
PI 548302	Japan	II	42.2	17.8	LP
PI 548311‡	Canada	0	42.0	20.4	LP
PI 548318‡	China	III	39.1	21.6	LP
PI 548325	Russia	0	41.5	19.7	LP
PI 548348	China	III	41.5	20.0	LP
PI 548352‡	North Korea	III	41.4	19	LP
PI 548356‡	North Korea	II	41.4	19.9	LP
PI 548360	North Korea	II	39.7	21.4	LP
PI 548362	United States	III	38.4	22.9	LP
PI 548379	China	0	38.4	20.9	LP
PI 548382‡	-	0	43.1	17.6	LP
PI 548391	China	II	43.0	20.3	LP
PI 548402‡	China	IV	38.2	18.5	LP
PI 548406	China	II	41.6	19.0	LP
PI 548438	North Korea	VI	44.7	19.2	LP
PI 548445	China	VII	45.7	19.0	LP
PI 548456	North Korea	VI	41.0	19.1	LP
PI 548461	United States	VIII	40.5	22.5	LP
PI 548477	United States	VI	42.9	20.2	LP
PI 548484	North Korea	VI	42.1	20.2	LP
PI 548485	China	VII	42.1	20.7	LP
PI 548488	China	V	43.8	18.9	LP
PI 548603	United States	IV	40.5	21.9	LP
PI 548657	United States	VII	40.3	21.9	LP
PI 71506	China	IV	41.0	22.6	LP
PI 80837‡	Japan	IV	42.4	18.2	LP
PI 88788	China	III	43.4	15.7	LP
Benning §	United States	VII	41.9	21.3	LP
Benning HP §	United States	VII	45.6	19.0	HP
Danbaekkong¶	South Korea	V	48.0	18.5	HP

Note: Protein and oil analyzed with near infrared spectroscopy NIR using a sample of approximately 200 seeds harvested in 2018.

† GSM1252 indicates the presence of the high protein allele (HP) or the low protein allele (LP).

‡ Protein and oil content retrieved from GRIN. https://npgsweb.ars-grin.gov/gringlobal/

§ Benning and Benning HP values are averages from three years of test (2019, 2020, and 2021).

¶ Danbaekkong value is average from two years of tests (2017 and 2021).

Table 2.3. Distribution of the high protein allele among the USDA *Glycine soja* core set as defined by La et al. (2019).

Name	Origin	MG	Protein (%)	Oil (%)	GSM1252[†]
PI 101404A	China	II	45.7	16.2	HP
PI 163453	China	VI	44.7	12.0	HP
PI 339871A	South Korea	V	42.9	16.6	HP
PI 342622A	Russia	I	43.7	16.1	HP
PI 366122	Japan	IV	44.1	16.6	HP
PI 378683	Japan	VI	46.7	16.4	HP
PI 378684B	Japan	VI	47.3	16.0	HP
PI 378686B	Japan	VI	46.0	16.3	HP
PI 378690	Japan	VII	45.3	16.3	HP
PI 378696B	Japan	VI	43.7	16.7	HP
PI 378697A	Japan	V	44.5	16.5	HP
PI 407020	Japan	V	44.0	16.8	HP
PI 407038	Japan	V	45.4	16.5	HP
PI 407042	Japan	V	44.9	16.3	HP
PI 407052	Japan	V	46.8	16.1	HP
PI 407059	Japan	-	46.7	16.1	HP
PI 407085	Japan	VI	44.8	16.5	HP
PI 407096	Japan	VII	47.2	16.3	HP
PI 407156	Japan	VI	44.7	16.5	HP
PI 407157	Japan	VI	47.8	16.3	HP
PI 407171	South Korea	IV	43.8	16.4	HP
PI 407179	South Korea	V	44.4	16.8	HP
PI 407191	South Korea	V	46.2	16.5	HP
PI 407195	South Korea	IV	44.4	16.5	HP
PI 407206	South Korea	V	46.4	16.3	HP
PI 407214	South Korea	V	46.7	16.4	HP
PI 407228	South Korea	V	49.5	15.8	HP
PI 407231	South Korea	V	44.4	16.5	HP
PI 407240	South Korea	V	46.3	16.5	HP
PI 407248	South Korea	V	44.6	16.6	HP
PI 407287	Japan	V	45.6	16.3	HP
PI 407300	China	V	46.1	16.2	HP
PI 407314	South Korea	V	44.2	16.9	HP
PI 424004B	South Korea	II	43.6	16.5	HP
PI 424007	South Korea	V	42.3	16.8	HP
PI 424025B	South Korea	V	46.3	16.4	HP
PI 424035	South Korea	V	43.3	16.7	HP
PI 424045	South Korea	V	42.6	16.5	HP
PI 424070B	South Korea	V	43.3	16.5	HP
PI 424082	South Korea	V	44.1	16.1	HP
PI 424083A	South Korea	V	45.4	16.4	HP
PI 424102A	South Korea	V	43.6	16.5	HP
PI 424116	South Korea	IV	43.7	16.6	HP
PI 424123	South Korea	V	44.0	16.1	HP
PI 447003A	China	0	43.8	16.8	HP

PI 458536	China	0	48.3	16.3	HP
PI 464890B	China	I	47.2	16.2	HP
PI 468916	China	III	44.0	10.1	HP
PI 479746B	China	II	46.6	16.1	HP
PI 479751	China	III	43.7	16.8	HP
PI 479752	China	I	41.2	16.4	HP
PI 479768	China	0	44.8	16.4	HP
PI 483466	China	V	43.9	16.2	HP
PI 507618	Japan	V	44.1	16.4	HP
PI 507624	Japan	VII	44.6	16.4	HP
PI 507641	Japan	V	45.9	16.6	HP
PI 507656	Japan	VII	45.9	16.3	HP
PI 507761	Russia	I	42.4	16.4	HP
PI 522209B	Russia	II	43.2	16.4	HP
PI 522226	Russia	0.00	43.3	16.3	HP
PI 522233	Russia	I	44.3	16.1	HP
PI 522235B	Russia	I	41.6	16.2	HP
PI 549032	China	III	44.0	15.9	HP
PI 549046	China	III	39.9	17.1	HP
PI 549048	China	III	41.0	17.6	HP
PI 562547	South Korea	V	41.2	16.5	HP
PI 562551	South Korea	V	43.9	16.5	HP
PI 562553	South Korea	V	47.4	16.3	HP
PI 562561	South Korea	V	47.1	16.0	HP
PI 562565	South Korea	IV	43.2	16.4	HP
PI 593983	Japan	III	45.0	16.7	HP
PI 597448D	China	0	45.2	16.2	HP
PI 597458C	China	V	43.5	17.2	HP
PI 597460A	China	IV	42.9	16.8	HP
PI 597461B	China	V	39.8	17.5	HP
PI 597462B	China	IV	42.5	17.1	HP
PI 639586	Russia	-	42.0	17.0	HP
PI 639588B	Russia	-	41.8	17.1	HP
PI 639621	Russia	-	41.8	17.1	HP
PI 639623A	Russia	-	44.2	16.5	HP
PI 639635	Russia	-	43.3	16.4	HP
Benning‡	United States	VII	41.9	21.3	LP
Benning HP‡	United States	VII	45.6	19.0	HP
Danbaekkong§	South Korea	V	48.0	18.5	HP

Note: Protein and oil contents for *G. soja* accessions were obtained from La et al., (2019).
Note: All *G. soja* accessions have black seed coat color.
† GSM1252 indicates the presence of the high protein allele (HP) or the low protein allele (LP).
‡ Benning and Benning HP values are averages from three years of test (2019, 2020, and 2021).
§ Danbaekkong value is average from two years of tests (2017 and 2021).

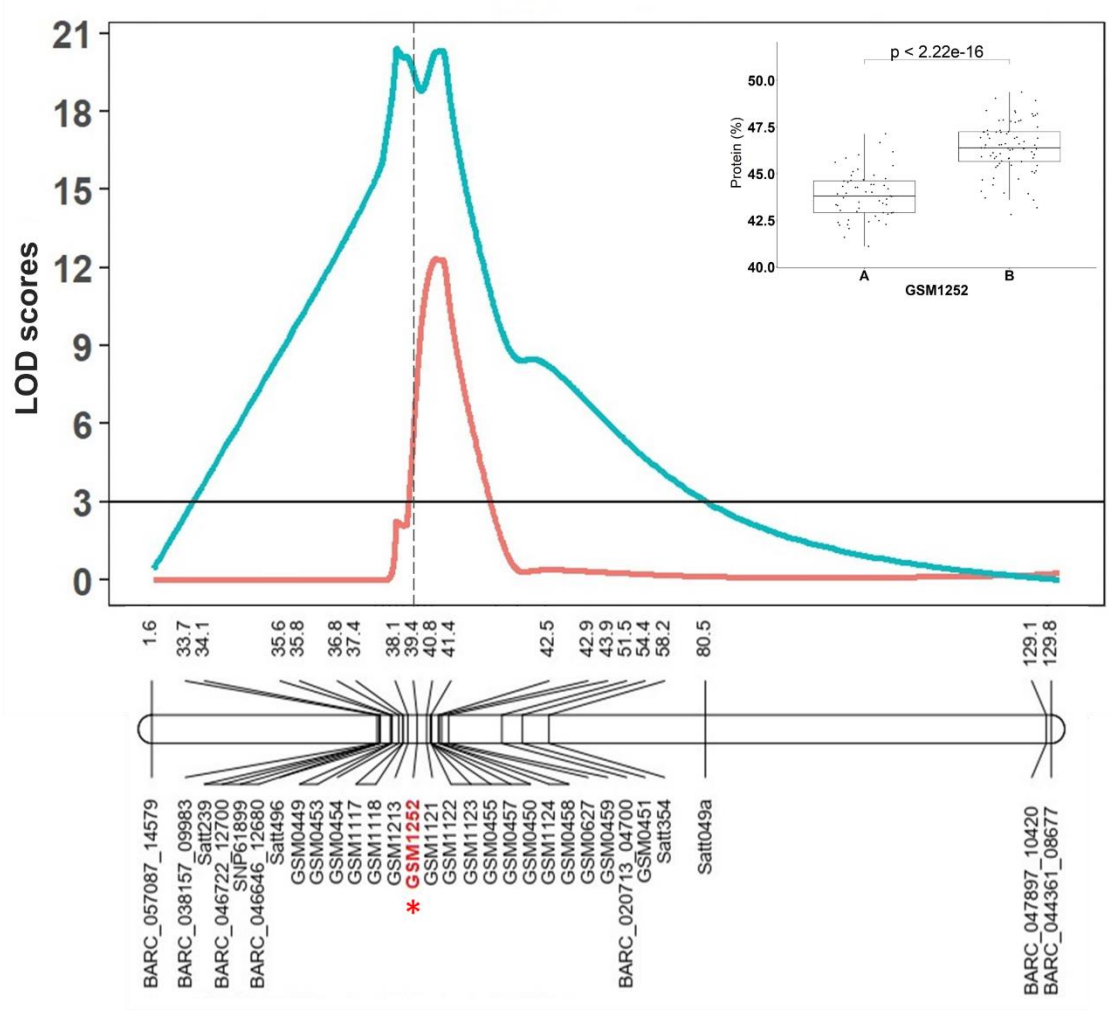

Figure 2.1. Chr 20 QTL region identified by Warrington et al. (2015) in the RIL population derived from Benning × Danbaekkong and saturated with additional SNPs and the gene specific marker GSM1252. Red lines indicate Composite Interval Mapping and blue lines indicate Simple Interval Mapping. Marker distances are given in centimorgan (cM).

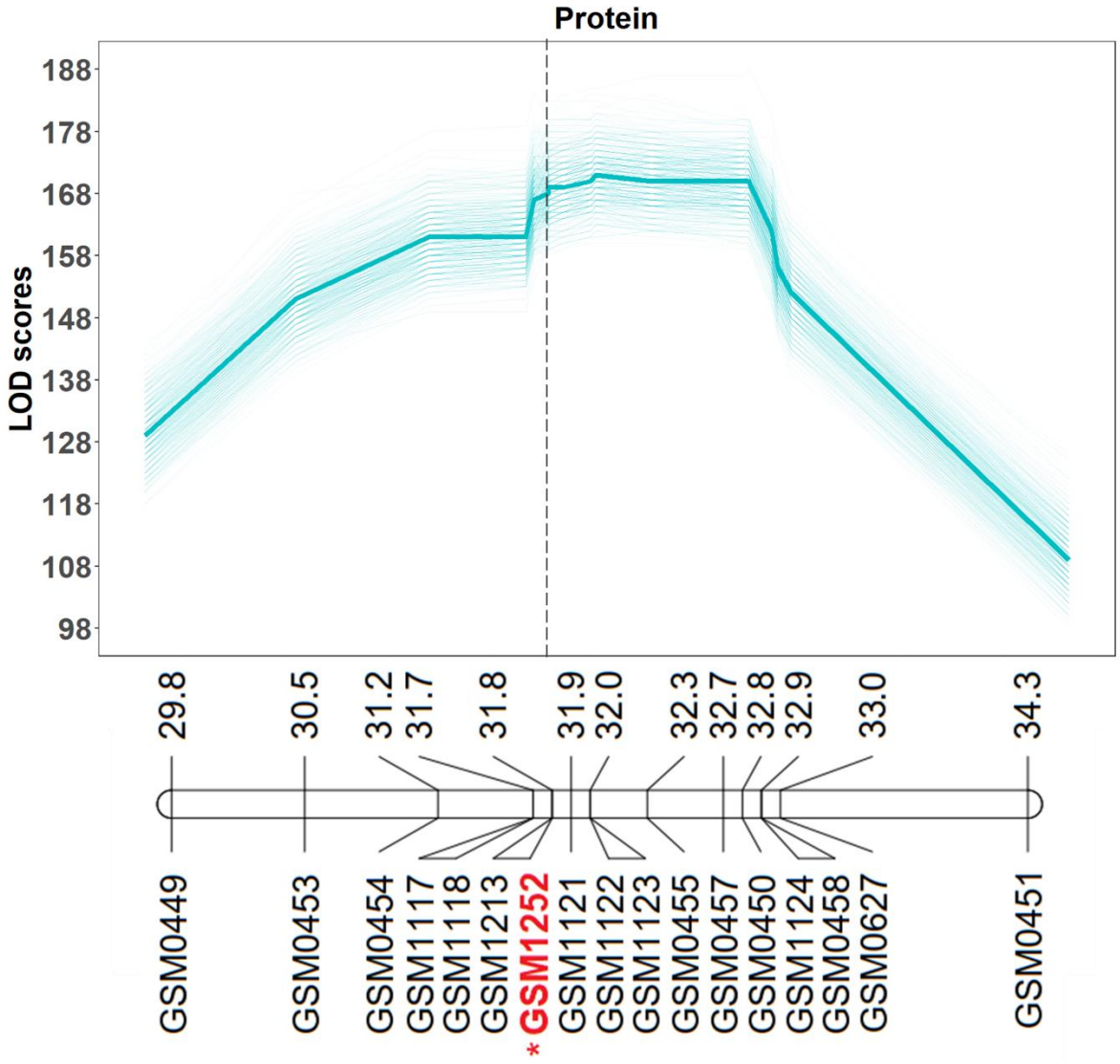

Figure 2.2. Multi-parent population QTL analysis for seed protein and oil content. QTL analysis was performed 500 times (5 random subsets with 100 replications) using the composite interval mapping function. The average LOD value of all values is indicated in the bold line.

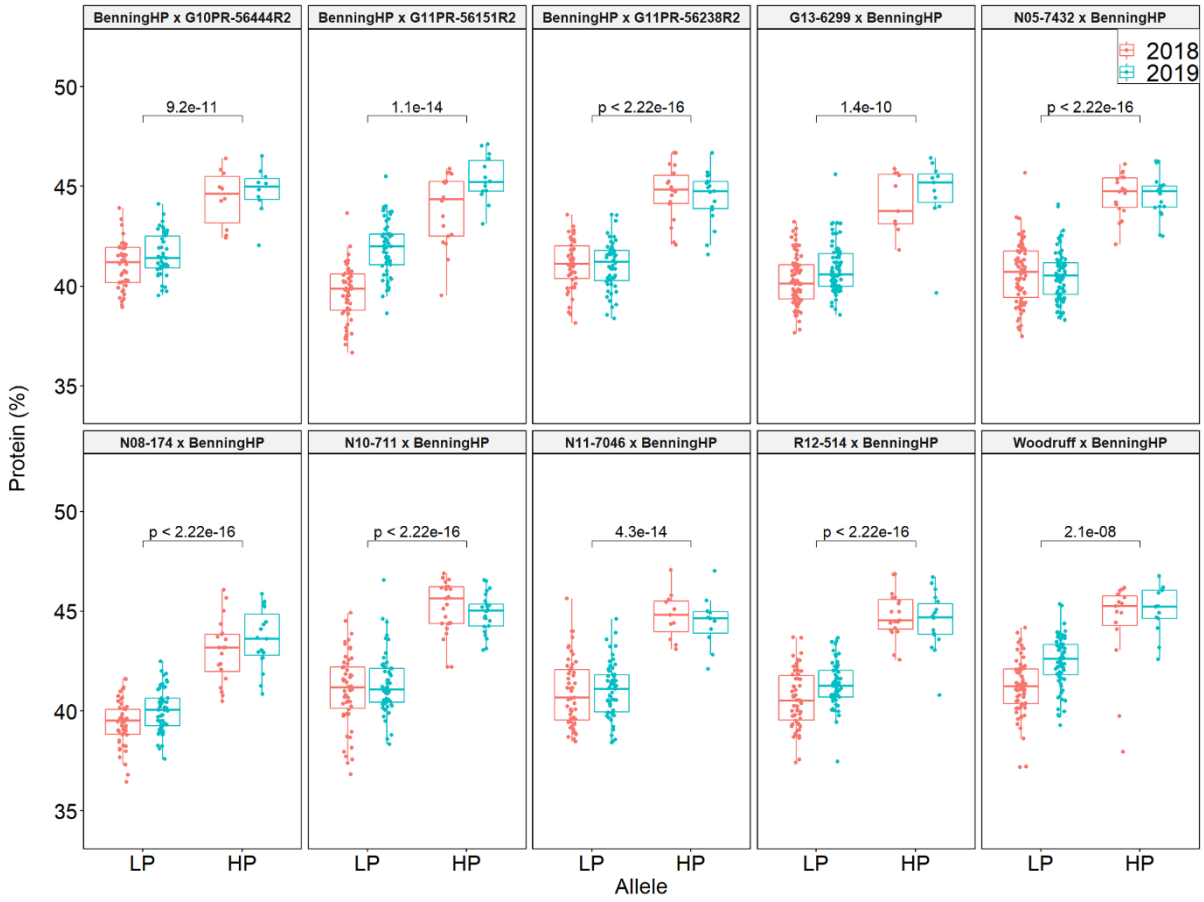

Figure 2.3. Effects of the Danbaekkong Chr 20 high protein allele on seed protein in 10 RIL populations evaluated in 2018-2019. X axis indicates the allele at the *Glyma.20g085100* (GSM1252). HP and LP represent the high and low protein alleles as indicated by the gene-specific marker GSM1252, respectively. Protein content is on the dry-matter basis.

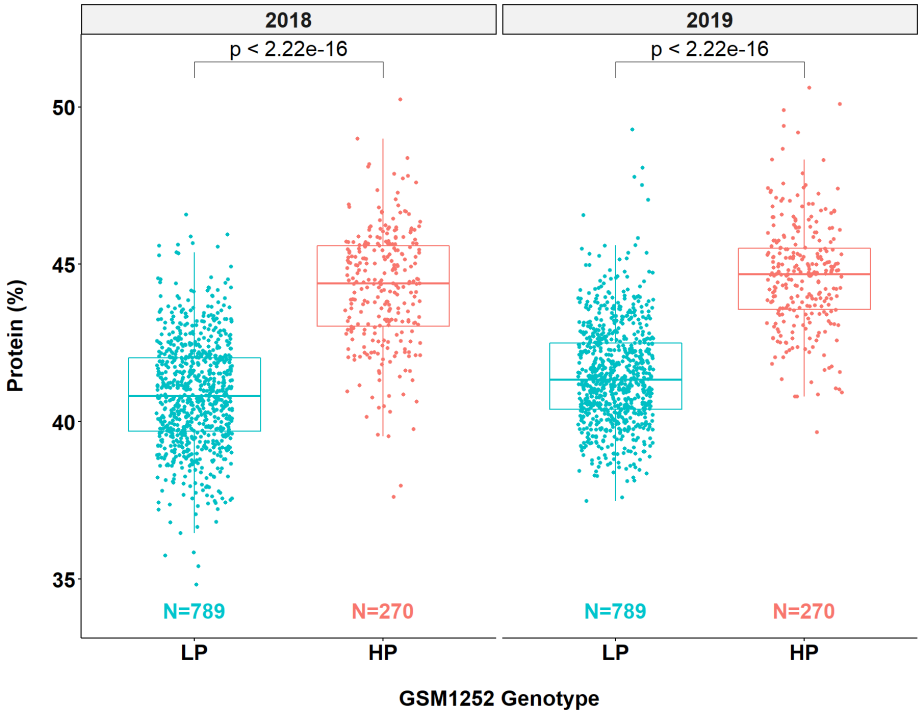

Figure 2.4. Effects of different alleles at *Glyma.20g085100* on protein content across the multi-parent RIL populations. HP indicates the high protein allele and LP indicates the low protein allele at GSM1252.

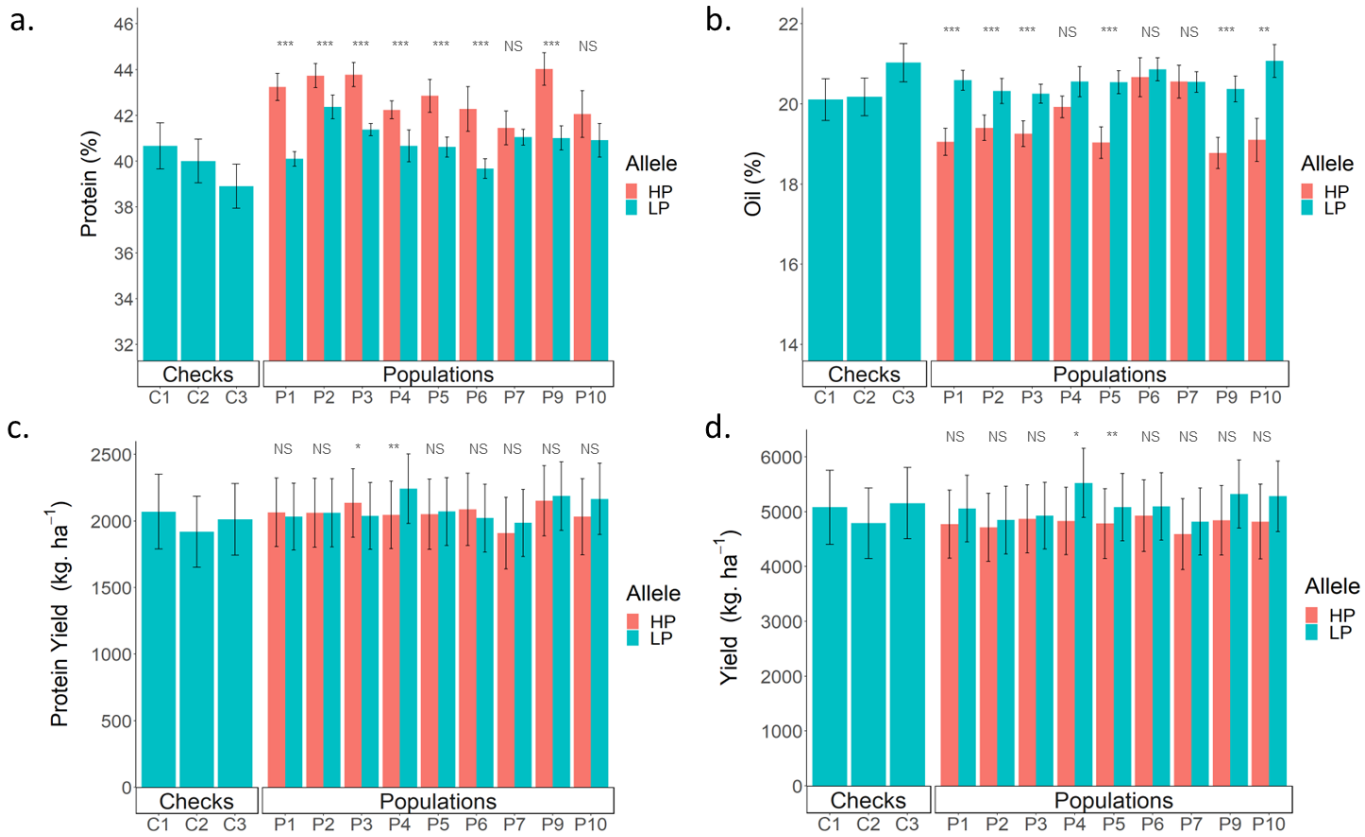

Figure 2.5. Comparison of lines with and without the Danbaekkong Chr 20 high protein allele in each population. Red indicates lines with the high protein allele (HP) and blue indicates lines with the low protein allele (LP). 5a) Comparison of the protein content, 5b) Oil content, 5c) Production of protein per hectare, and 5d) Seed yield. One hundred and three RILS were evaluated in five environments (Athens, Plains, and Tifton, GA). *, ** and *** Indicates significance at the 0.05, 0.01, and 0.001 probability level and NS indicates not significant.

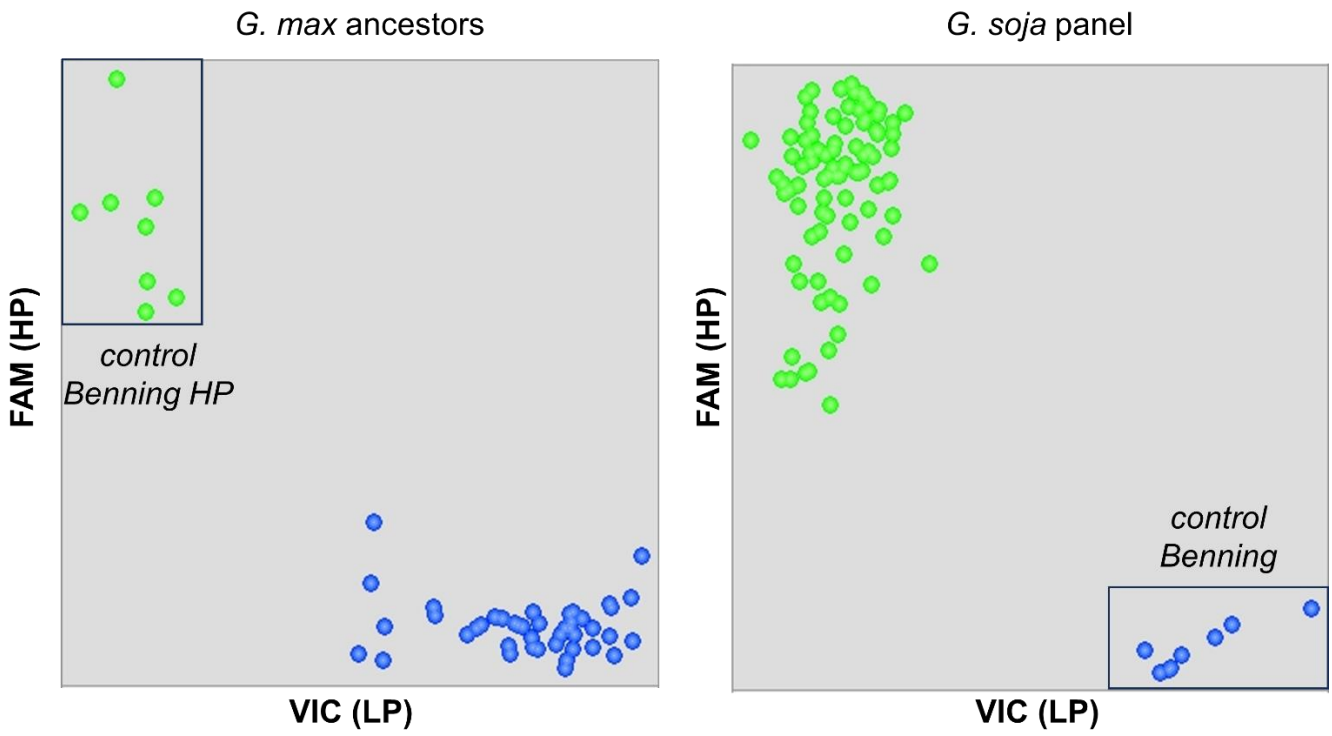

Figure 2.6. Genotyping results of 35 North America *Glycine max* ancestors and 79 *Glycine soja* accessions with a gene specific marker GSM1252. Benning and Benning HP were used as controls for the low protein (LP) and the high protein allele (HP), respectively.

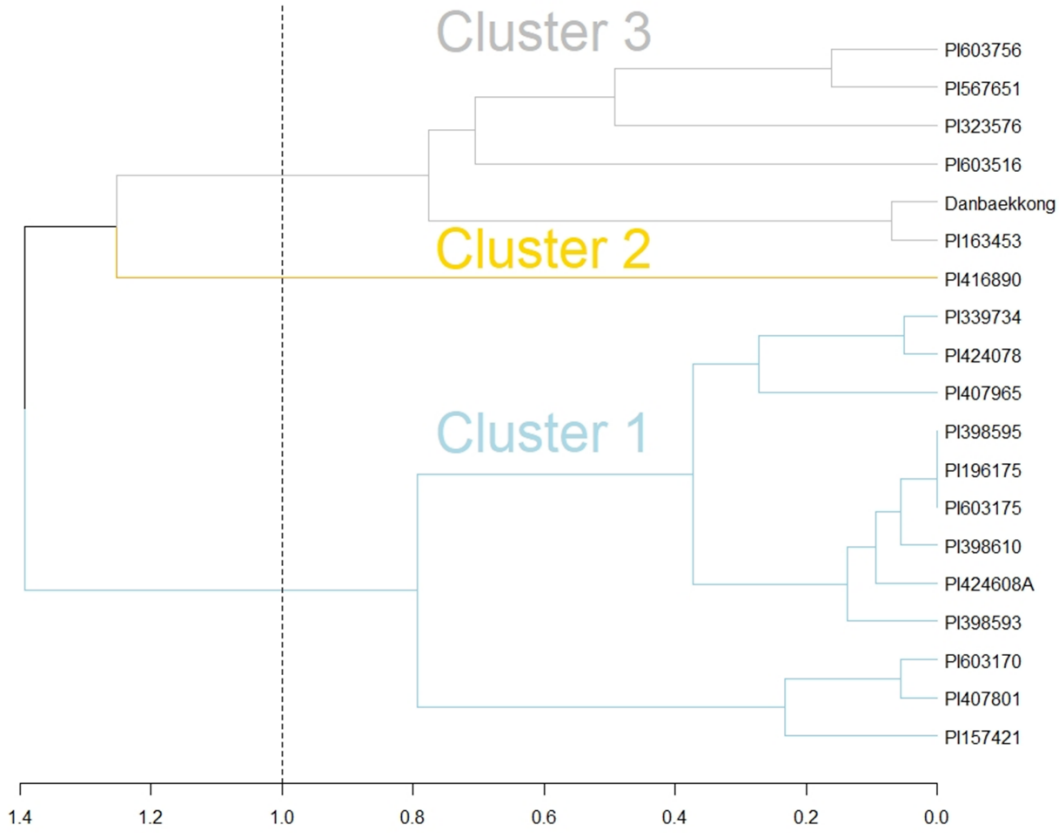

Figure 2.7. Comparison of PI 163453 and Danbaekkong haplotypes with the three introgression groups identified by Goettel et al. (2023) using hierarchical complete linkage cluster analysis. Analysis was based on 82 SNPs from the SoySNP50K at the Chr 20 QTL region between 29 and 34 Mb.

Table 2.S1. Sequenced accessions information.

Accession	Name	Type	Origin	Haplotype‡	Protein (%)
PI 595645	Benning	Cultivar	United States	-	42.2
PI 619083	Danbaekkong	Cultivar	South Korea	+	51.5
PI 163453	Quail Haven	*G. soja*	China	+	44.7*
PI 398589	KAS 390-3	Landrace	South Korea	+	55.1*
PI 408012	KAERI 548-5	Landrace	South Korea	+	51.3*
PI 602447	BARC-14 nodulated	Cultivar	United States	+	47.6*
PI 468916†		*G. soja*	China	+	44.0*

† Sequencing data retrieved from https://www.ncbi.nlm.nih.gov/sra/?term=SRP045129

‡ Haplotype in the QTL region. (-) Low protein haplotype, (+) High protein haplotype.

*Values obtained from USDA GRIN at https://npgsweb.ars-grin.gov/gringlobal/search

Table 2.S2. Protein and oil content of progenitors and parents of the 10 populations.

Line	Protein (%)†	Oil (%)†	Derived RILs
Danbaekkong‡	48.0	18.5	NA
Benning HP	45.6	19.0	NA
Benning	41.9	21.3	NA
R12-514	43.6	19.8	119
G11PR-56151R2	42.3	20.0	119
Woodruff	42.1	19.3	115
G11PR-56238R2	41.4	19.8	109
N10-711	41.3	19.3	114
G10PR-56444R2	40.5	21.0	105
N05-7432	39.9	20.9	121
G13-6299	39.7	20.7	111
N08-174	39.0	21.9	102
N11-7046	37.8	21.4	100
Total			1115

† Protein and oil are averages of three years (2019, 2020, and 2021).
‡ Protein and oil are averages of two years (2017 and 2021).

Table 2.S3. Molecular markers used to fine map the Chr 20 QTL in the Benning × Danbaekkong population.

Marker Name	Type	dbSNP ID	Wm82.a2.v1 Coordinate	Alleles	Reference
BARC_057077_14568	SNP	ss107925764	14507618	A_G	Warrington et al., 2015
BARC_038157_09983	SNP	ss107919194	24536596	A_G	Warrington et al., 2015
Satt239	SSR	-	25275083	(AAT)22	Warrington et al., 2015
BARC_046722_12700	SNP	ss107920968	25446707	C_T	Warrington et al., 2015
GSM0012	SNP	-	26201368	G_T	Warrington et al., 2015
Satt496	SSR	-	27664504	(ATT)13	Warrington et al., 2015
GSM0449	SNP	ss715637188	29813037	T_C	New Marker
GSM0453	SNP	ss715637217	30546685	T_C	New Marker
GSM0454	SNP	ss715637241	31195048	G_T	New Marker
GSM1117	SNP	-	31666190	T_A	New Marker
GSM1118	SNP	-	31707347	T_C	New Marker
GSM1213	SNP	-	31775201	A_G	New Marker
GSM1252	Insertion	-	31778817	Insertion+/-	New Marker
GSM1121	SNP	-	31851276	A_G	New Marker
GSM1122	SNP	-	31981547	T_C	New Marker
GSM1123	SNP	-	32010590	G_A	New Marker
GSM0455	SNP	ss715637294	32282623	T_C	New Marker
GSM0457	SNP	ss715637315	32721955	A_G	New Marker
GSM0450	SNP	ss715637316	32752215	T_C	New Marker
GSM1124	SNP	-	32865400	C_A	New Marker
GSM0458	SNP	ss715637323	32894943	T_C	New Marker
GSM0627	Insertion	-	32958630	-	New Marker
BARC_020713_04700	SNP	ss107916725	34052339	C_T	Warrington et al., 2015
GSM0451	SNP	ss715637431	34314607	G_A	New Marker
Satt354	SSR		34569176		Warrington et al., 2015
Satt049	SSR		36842373	(AAT)16	Warrington et al., 2015

Table 2.S4. Markers used to dissect the Chr 20 QTL across the multi-parent populations.

Marker Name	SNP ID	Position Wm82.a2 (bp)	Alleles	Favorable allele	Type	Genes
GSM0449	ss715637188	29813037	T_C	C	Intergenic	*Glyma.20g080000* *Glyma.20g080100*
GSM0453	ss715637217	30546685	T_C	C	Intergenic	*Glyma.20g081200* *Glyma.20g081300*
GSM0454	ss715637241	31195048	G_T	T	Intergenic	*Glyma.20g082700* *Glyma.20g082800*
GSM1117	-	31666190	T_A	A	Intragenic	*Glyma.20g084500*
GSM1118	-	31707347	T_C	C	Intragenic	*Glyma.20g084900*
GSM1213	-	31775201	A_G	G	Intragenic	*Glyma.20g085100*
GSM1252	-	31778817	Insertion+/-	-	Intragenic	*Glyma.20g085100*
GSM1121	-	31851276	A_G	G	Intragenic	*Glyma.20g085500*
GSM1122	-	31981547	T_C	C	Intragenic	*Glyma.20g085700*
GSM1123	-	32010590	G_A	A	Intragenic	*Glyma.20g086000*
GSM0455	ss715637294	32282623	T_C	C	Intergenic	*Glyma.20g086700* *Glyma.20g086800*
GSM0457	ss715637315	32721955	A_G	G	Intergenic	*Glyma.20g087500* *Glyma.20g087600*
GSM0450	ss715637316	32752215	T_C	C	Intergenic	*Glyma.20g087600* *Glyma.20g087700*
GSM1124	-	32865400	C_A	A	Intragenic	*Glyma.20g087800*
GSM0458	ss715637323	32894943	T_C	C	Intergenic	*Glyma.20g087900* *Glyma.20g088000*
GSM0627	-	32958630	Insertion+/-	-	Intragenic	*Glyma.20g088500*
GSM0451	ss715637431	34314607	G_A	A	Intragenic	*Glyma.20g099900*

Table 2.S5. Sequence of the markers developed to saturate the Chr 20 QTL

Marker	Primer Sequence
GSM0449	Fam: GAAGGTGACCAAGTTCATGCTTCGCTGACTCTGCCACTGc
	Hex: GAAGGTCGGAGTCAACGGATTTCGCTGACTCTGCCACTGt
	Rev: ATGGACGACGGAGTAAGCAT
GSM0450	Fam: GAAGGTGACCAAGTTCATGCTGGAGCAGAAGAGGGGGATGc
	Hex: GAAGGTCGGAGTCAACGGATTGGAGCAGAAGAGGGGGATGt
	Rev: TGCTGGAACCTGGACGAT
GSM0451	Fam: GAAGGTGACCAAGTTCATGCTCGTTGAGTGACTGAGAGCCCAa
	Hex: GAAGGTCGGAGTCAACGGATTCGTTGAGTGACTGAGAGCCCAg
	Rev: GCCCTATACTTACAGCAAAGAAGCA
GSM0453	Fam: GAAGGTGACCAAGTTCATGCTGTCACCACTACCGACATTATCGc
	Hex: GAAGGTCGGAGTCAACGGATTGTCACCACTACCGACATTATCGt
	Rev: TTCAAACAAAGCCAGAAATGC
GSM0454	Fam: GAAGGTGACCAAGTTCATGCTGAGCAAAAGAGAGGGAATCAg
	Hex: GAAGGTCGGAGTCAACGGATTGAGCAAAAGAGAGGGAATCAt
	Rev: GCTGACGAGAACTTGGGATG
GSM0455	Fam: GAAGGTGACCAAGTTCATGCTCAACCTTCTTCTTCTACTTCTATCc
	Hex: GAAGGTCGGAGTCAACGGATTCAACCTTCTTCTTCTACTTCTATCt
	Rev: TGTTGCTCATGCTAAGCCATA
GSM0457	Fam: GAAGGTGACCAAGTTCATGCTTTGTGGCTATTGAGAGTAACa
	Hex: GAAGGTCGGAGTCAACGGATTTTGTGGCTATTGAGAGTAACg
	Rev: GCTACTGCTCTTTCTTCATTTACGC
GSM0458	Fam: GAAGGTGACCAAGTTCATGCTACAATGGGTGAAGTGAAGc
	Hex: GAAGGTCGGAGTCAACGGATTACAATGGGTGAAGTGAAGt
	Rev: GTAACCAGCGAGTACATGACCAA
GSM0627	Fam: GAAGGTGACCAAGTTCATGCTACAGATTTACACAGTACGTTAAGGACAGT
	Hex: GAAGGTCGGAGTCAACGGATTCATTCATCACCCAAAAGTACGTTAAG
	Rev: AGTGGTCAAGAGGAAAACTTGTGAA
GSM1117	Fam: GAAGGTGACCAAGTTCATGCTCAGTTCCATTCAGATTTCACATTTGCa
	Hex: GAAGGTCGGAGTCAACGGATTCAGTTCCATTCAGATTTCACATTTGCt
	Rev: GCTTGTTAGTTTGTCATCCCTTTCAT
GSM1118	Fam: GAAGGTGACCAAGTTCATGCTAACCGAAGAAGAGCCACCCAc
	Hex: GAAGGTCGGAGTCAACGGATTCAACCGAAGAAGAGCCACCCAt
	Rev: CTTTGTTGGTCCAGTTCTTCGCTATTAG
GSM1121	Fam: GAAGGTGACCAAGTTCATGCTAGCACAAGGAGATCAAATTAAGAACCg
	Hex: GAAGGTCGGAGTCAACGGATTTAAGCACAAGGAGATCAAATTAAGAACCa

	Rev: TGAGTAGCTGTATAGTTCAAATTGCTTG
GSM1122	Fam: GAAGGTGACCAAGTTCATGCTCCTCCAATTGCAGACGTAACACc Hex: GAAGGTCGGAGTCAACGGATTCCTCCAATTGCAGACGTAACACt Rev: TTCCCTTTCTAGGAATGAGGAAGAACTA
GSM1123	Fam: GAAGGTGACCAAGTTCATGCTTTCTCTGAATCAACAACACAGGAAATTa Hex: GAAGGTCGGAGTCAACGGATTTTCTCTGAATCAACAACACAGGAAATTg Rev: CCACATACATGGGTGATGAGAATAAC
GSM1124	Fam: GAAGGTGACCAAGTTCATGCTGGAAAAATAATCTAAGCCTCGGTCAa Hex: GAAGGTCGGAGTCAACGGATTGGAAAAATAATCTAAGCCTCGGTCAc Rev: GGTGGGAGTTGAGGTTAAGGG
GSM1213*	Fam: GAAGGTGACCAAGTTCATGCTCATTAACACTAAATATACATGATCGAGAc Hex: GAAGGTCGGAGTCAACGGATTCATTAACACTAAATATACATGATCGAGAt Rev: CACCATGTTGCAGGATGTTG
GSM1252	Fwd: CCTTGTTTATGGCTCTCTCC Rev: TGCATCAACCAAGCCTTAT Probe FAM (INS-): GCGGCAAGCATACTGCATTTT Probe VIC (INS+): CGGCAAGCATACAACAACAACA

* Marker designed in reverse 3`→ 5`

Table 2.S6. Protein and oil content of 1115 RILs with the high protein allele (HP) and RILs with the low protein allele (LP) (from 10 populations evaluated under field conditions in 2018 and 2019 in Athens, Georgia.

Population ID	Pedigree	Protein (%)			Oil (%)		
		HP	LP	Difference	HP	LP	Difference
P1	G13-6299 × Benning HP	44.5	40.8	3.7***	19.7	21.4	-1.6***
P2	Woodruff × Benning HP	45.4	42.1	3.2***	19.3	21.2	-1.9***
P3	N10-711 × Benning HP	45.2	41.8	3.4***	19.1	21.1	-2.0***
P4	N05-7432 × Benning HP	44.3	40.7	3.6***	19.9	21.6	-1.7***
P5	N11-7046 × Benning HP	44.2	41.5	2.7***	19.5	21.2	-1.8***
P6	N08-174 × Benning HP	43.5	39.9	3.6***	20.1	22.0	-1.9***
P7	R12-514 × Benning HP	44.4	41.2	3.2***	19.9	21.6	-1.7***
P8	Benning HP × G10PR-56444R2	44.0	41.4	2.6***	20.0	21.4	-1.4***
P9	Benning HP × G11PR-56151R2	44.5	41.0	3.5***	19.9	21.8	-1.9***
P10	Benning HP × G11PR-56238R2	44.4	41.2	3.1***	19.3	21.3	-2.0***
Average		44.5	41.1	3.3***	19.7	21.5	-1. 8***

*, **, *** Significant difference at the 0.05, 0.01 and 0.001 probability level in a t-test, respectively.

Table 2.S7. Comparison of lines with the high protein allele (HP) and lines with the low protein allele (LP) in each pedigree. One hundred and three Recombinant Inbred Lines (RILs) were evaluated in yield trials from five environments†.

Population ID	Pedigree	N		Yield (kg ha^{-1})		Oil (%)		Protein (%)		Protein yield (kg ha^{-1})		Difference (HP-LP)			
		HP	LP	HP	LP	HP	LP	HP	LP	HP	LP	Yield (kg ha^{-1})	Oil (%)	Protein (%)	Protein yield (kg ha^{-1})
P1	G13-6299 × Benning HP	3	14	4764	5044	19.1	20.6	43.2	40.1	2060	2022	-279	-1.5	3.1	38
P2	Woodruff × Benning HP	4	4	4691	4841	19.4	20.3	43.7	42.4	2051	2051	-150	-0.9	1.4	0
P3	N10-711 × Benning HP	4	23	4855	4910	19.2	20.3	43.8	41.4	2125	2031	-55	-1.0	2.4	94
P4	N05-7432 × Benning HP	8	2	4811	5529	19.9	20.5	42.2	40.7	2031	2249	-719	-0.6	1.5	-218
P5	N11-7046 × Benning HP	2	6	4767	5069	19	20.5	42.8	40.6	2042	2059	-302	-1.5	2.2	-17
P6	N08-174 × Benning HP	1	6	4934	5079	20.7	20.9	42.3	39.7	2086	2015	-145	-0.2	2.6	72
P7	R12-514 × Benning HP	2	11	4566	4807	20.6	20.5	41.4	41.0	1892	1973	-241	0.0	0.4	-81
P8	Benning HP × G10PR-56444R2	NA	4	NA	5311	NA	20.8	NA	40.9	NA	2173	NA	NA	NA	NA
P9	Benning HP × G11PR-56151R2	2	4	4841	5305	18.8	20.4	44.0	41.0	2131	2175	-464	-1.6	3.0	-44
P10	Benning HP × G11PR-56238R2	1	2	4796	5257	19.1	21.1	42.0	40.9	2016	2150	-461	-2.0	1.1	-134
	Average			4781	5115	19.5	20.6	42.8	40.9	2048	2080	-313**	-1.0**	2.0**	-32.0^{NS}

**Statistically different in the paired t-test, $p<0.01$

† Athens-2020, Plains-2020, Athens-2021, Plains-2021, Tifton-2021.

Table 2.S8. Performance of selected RILs across five environments†. Line performance was compared to the highest yielding check. HP indicates the presence of the high protein allele and LP is the low protein allele.

Line ID	Pedigree	GSM1252	Yield (kg ha^{-1})	% Check yield	Protein (%)	Protein yield (kg ha^{-1})
AGS 738RR	Commercial check	LP	5160	100	38.91	2020
G19-11395	N05-7432 × Benning HP	LP	5880	113.8	40.6	2380
G19-2050R2	Benning HP × G10PR-56444R2	LP	5810	112.6	41.2	2400
G19-11112	G13-6299 × Benning HP	LP	5570	107.8	39.1	2170
G19-2192R2	Benning HP × G11PR-56151R2	LP	5540	107.3	41.1	2280
G19-2308R2	Benning HP × G11PR-56238R2	LP	5530	107.0	40.6	2250
G19-11120	G13-6299 × Benning HP	LP	5450	105.5	40.7	2220
G19-2003R2	Benning HP × G10PR-56444R2	LP	5420	105.0	40.4	2200
G19-11535	N11-7046 × Benning HP	LP	5420	104.9	42.3	2300
G19-2115R2	Benning HP × G11PR-56151R2	LP	5420	104.9	40.4	2190
G19-11605	N08-174 × Benning HP	LP	5410	104.8	39.1	2110
G19-11114	G13-6299 × Benning HP	LP	5410	104.7	39.8	2160
G19-11029	G13-6299 × Benning HP	LP	5390	104.3	39.6	2140
G19-11507	N11-7046 × Benning HP	LP	5380	104.2	40.6	2190
G19-11257	N10-711 × Benning HP	LP	5370	104.0	42.7	2290
G19-2229R2	Benning HP × G11PR-56151R2	LP	5360	103.9	41.7	2240
G19-11204	Woodruff × Benning HP	LP	5320	103.1	42.7	2280
G19-11637	N08-174 × Benning HP	LP	5240	101.6	39.0	2050
G19-11035	G13-6299 × Benning HP	LP	5210	100.8	39.6	2070
G19-11462	N05-7432 × Benning HP	LP	5190	100.5	40.7	2110
G19-11191	Woodruff × Benning HP	HP	5190	100.5	43.6	2280

† Athens-2020, Plains-2020, Athens-2021, Plains-2021, Tifton-2021.

Table 2.S9. Analysis of the presence of the 321 bp insertion in *Glyma.20g085100* in 35 *Glycine soja* accessions based on genome sequencing data.

PI number/cultivar name	Species	MG	Origin	321 bp insertion
Benning	*G. max*	VII	United States	Yes
Danbaekkong	*G. max*	V	South Korea	No
PI 163453	*G. soja*	VI	China	No
PI 468916	*G. soja*	III	China	No
PI 378699A	*G. soja*	VII	Japan	No
PI 407038	*G. soja*	V	Japan	No
PI 407085	*G. soja*	VI	Japan	No
PI 407156	*G. soja*	VI	Japan	No
PI 407175	*G. soja*	IV	South Korea	No
PI 407179	*G. soja*	V	South Korea	No
PI 407183	*G. soja*	V	South Korea	No
PI 407191	*G. soja*	V	South Korea	No
PI 407229	*G. soja*	V	South Korea	No
PI 407231	*G. soja*	V	South Korea	No
PI 407240	*G. soja*	V	South Korea	No
PI 407262	*G. soja*	VI	South Korea	No
PI 407270	*G. soja*	VI	South Korea	No
PI 407287	*G. soja*	V	Japan	No
PI 407307	*G. soja*	VI	China	No
PI 407315	*G. soja*	V	South Korea	No
PI 407318A	*G. soja*	V	South Korea	No
PI 424014	*G. soja*	V	South Korea	No
PI 424070B	*G. soja*	V	South Korea	No
PI 424107A	*G. soja*	VI	South Korea	No
PI 483466	*G. soja*	V	China	No
PI 507609	*G. soja*	VI	Japan	No
PI 507615	*G. soja*	VI	Japan	No
PI 507619B	*G. soja*	VI	Japan	No
PI 507624	*G. soja*	VII	Japan	No
PI 507638	*G. soja*	VI	Japan	No
PI 507667	*G. soja*	VI	Japan	No
PI 507830B	*G. soja*	0	Russia	No
PI 507847	*G. soja*	II	Russia	No
PI 532450	*G. soja*	I	China	No
PI 549047	*G. soja*	III	China	No
PI 562534	*G. soja*	NA	South Korea	No
PI 562557	*G. soja*	NA	South Korea	No
PI 562568	*G. soja*	NA	South Korea	No
PI 639623A	*G. soja*	0	Russia	No

Table 2.S10. Genetic similarity between Benning, Danbaekkong, PI 163453 and PI 468916 at the Chr 20 locus (30 – 34 Mb, Wm82.a2.v1) calculated with 6,353 SNPs.

Line ID	Benning	Danbaekkong	PI 163453	PI 468916
Benning	1.00			
Danbaekkong	0.53	1.00		
PI 163453	0.53	0.99	1.00	
PI 468916	0.46	0.49	0.49	1.00

Table 2.S11. Accessions in the USDA Germplasm Collection with the highest similarity to Danbaekkong and genotyping result with the marker GSM1252. HP is the high protein allele and LP is the low protein allele at *Glyma.20g085100*.

Line Name	Similarity†	Species	Origin	GSM1252	Protein (%)‡	Oil (%)‡
Danbaekkong	1	*G. max*	South Korea	HP	45.43	18.99
BARC-6	1	*G. max*	United States	HP	52.21	17.56
BARC-14 nodulated	0.99	*G. max*	United States	HP	48.21	18.31
Qi du qing pi dou	0.99	*G. max*	China	HP	46.04	17.53
Kwangankong	0.95	*G. max*	South Korea	HP	44.18	19.24
D76-8070	0.95	*G. max*	United States	HP	49.29	18.07
PI 163453	0.94	*G. soja*	China	HP	44.62	14.85
Benning		*G. max*	United States	LP	42.02	22.24
Benning HP		*G. max*	United States	HP	43.32	21.21

† Based on 92 SNPs from the SoySNP50K located between 30 and 34 Mb on chromosome 20.

‡ Protein and oil content evaluated in a sample of 200 seeds from plants grown in greenhouse in 2022

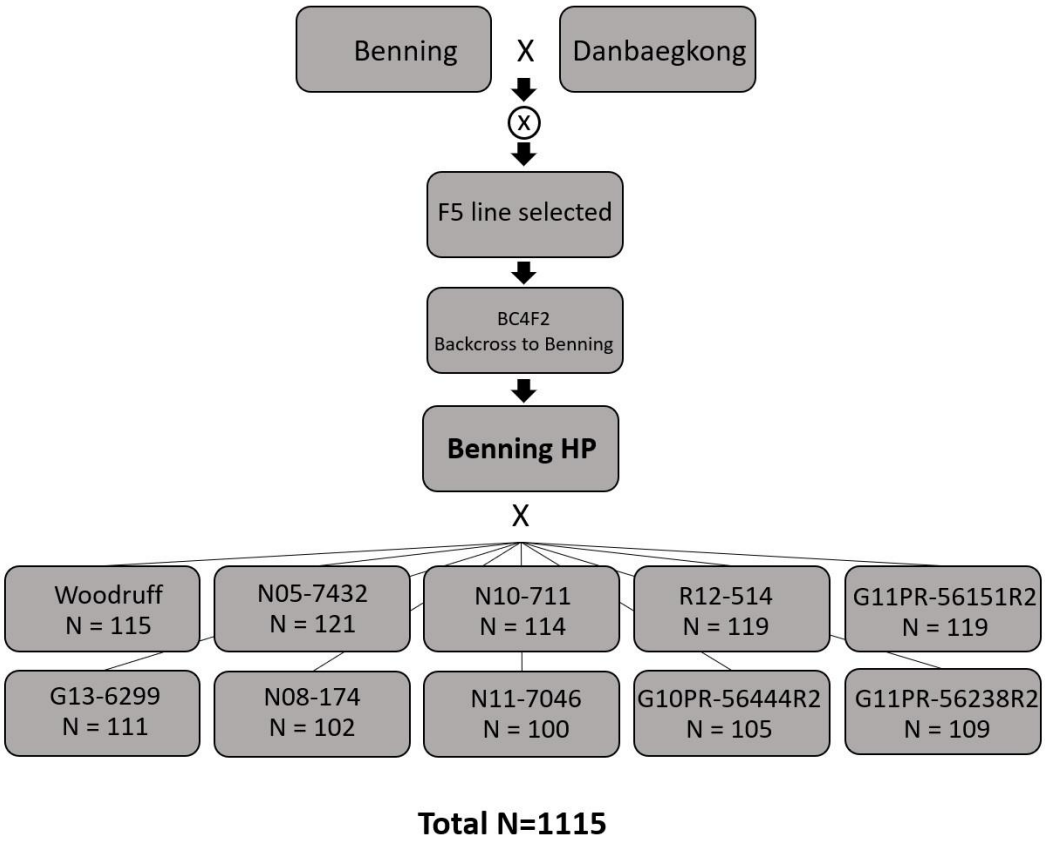

Figure 2.S1. Flowchart depicting the development of Benning HP and the derived RIL populations.

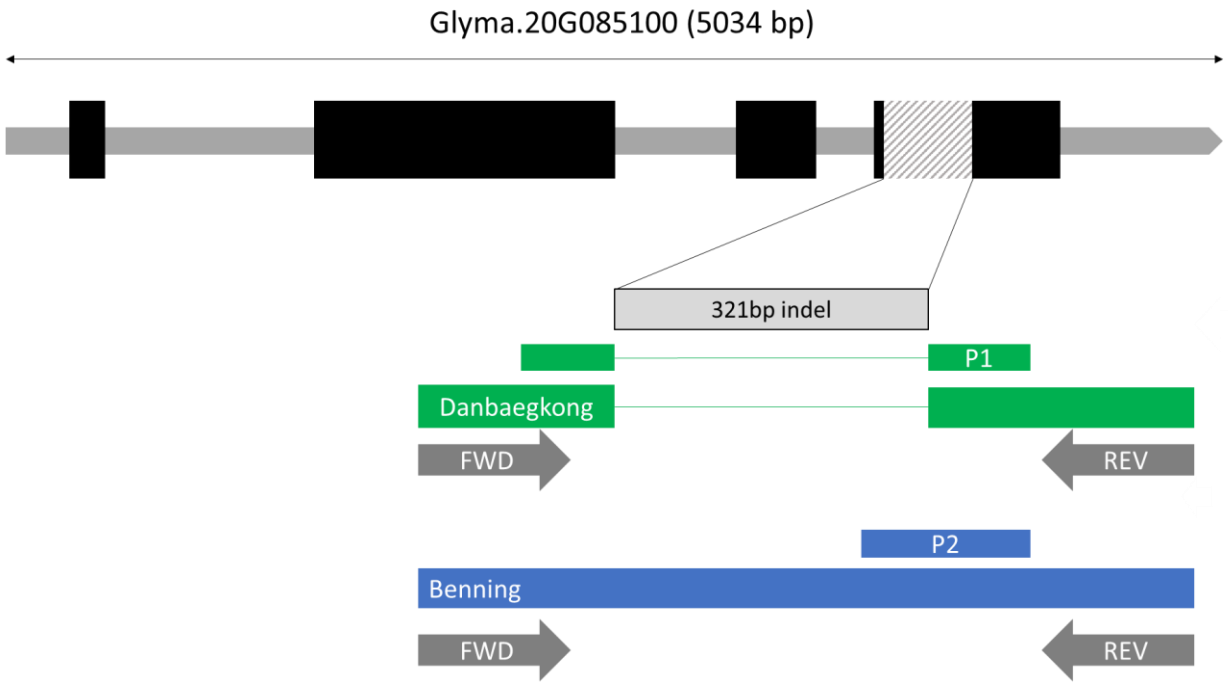

Figure 2.S2. Design of TaqMan marker GSM1252. P1 and P2 indicate the DNA probes and gray arrows indicate the primers.

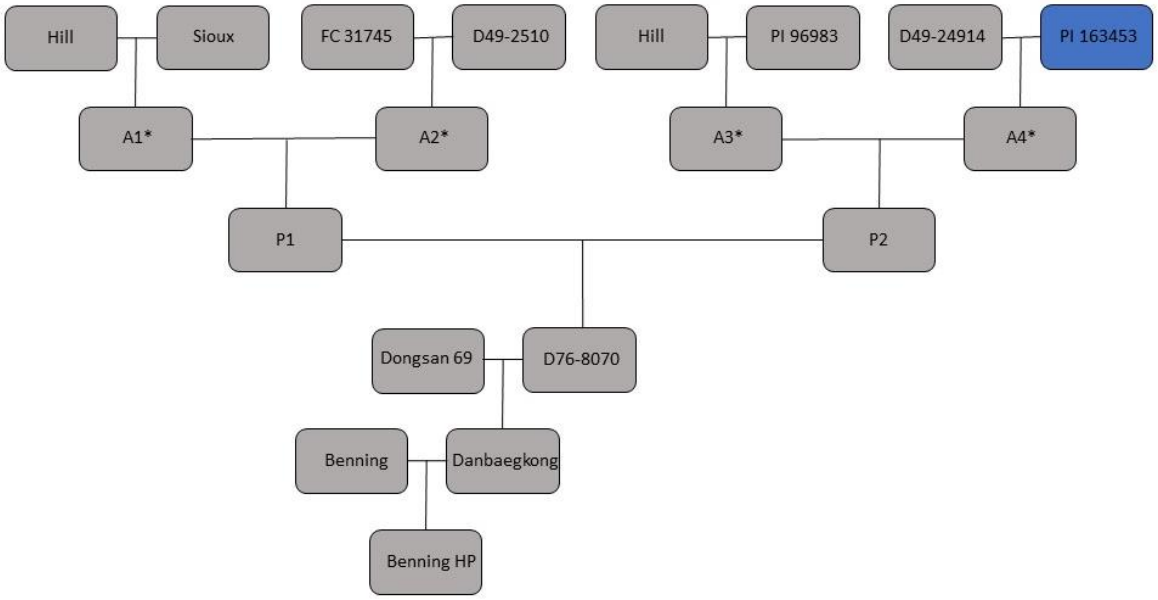

Figure 2.S3. Danbaekkong ancestry. * Indicates ancestor lines selected for resistance to bacterial pustule, shattering, and protein content higher than 45% (Hartwig, 1990).

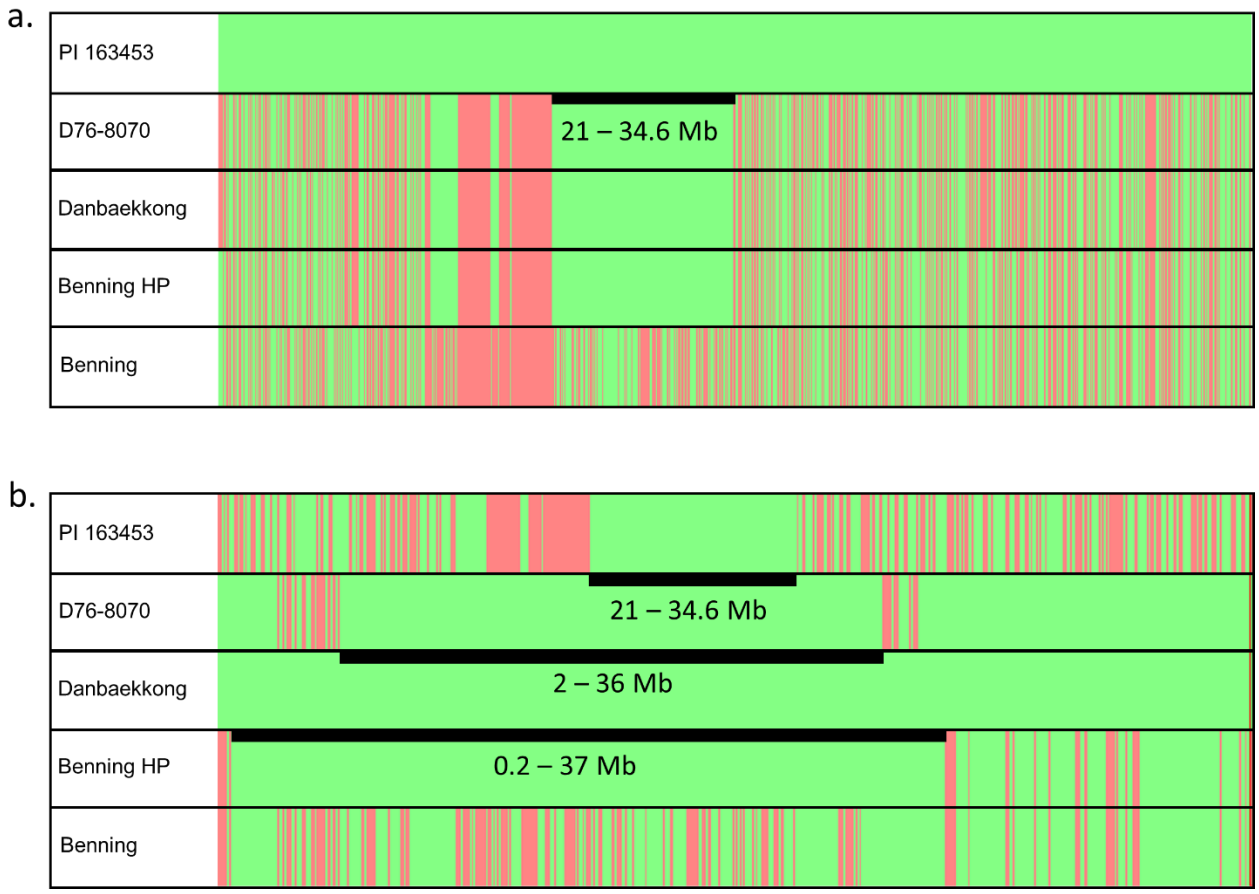

Figure 2.S4. PI 163453 Chr 20 fragment transferred to D76-8070 and subsequently to Danbaekkong and Benning HP. Comparison based on 1316 SNPs from the Soy50KSNP data. A) comparison using PI 163463 as the reference. B) comparison using Danbaekkong as the reference. Green indicates single nucleotide polymorphisms (SNPs) matching the reference genotype, and red denotes SNPs not matching the reference genotype.

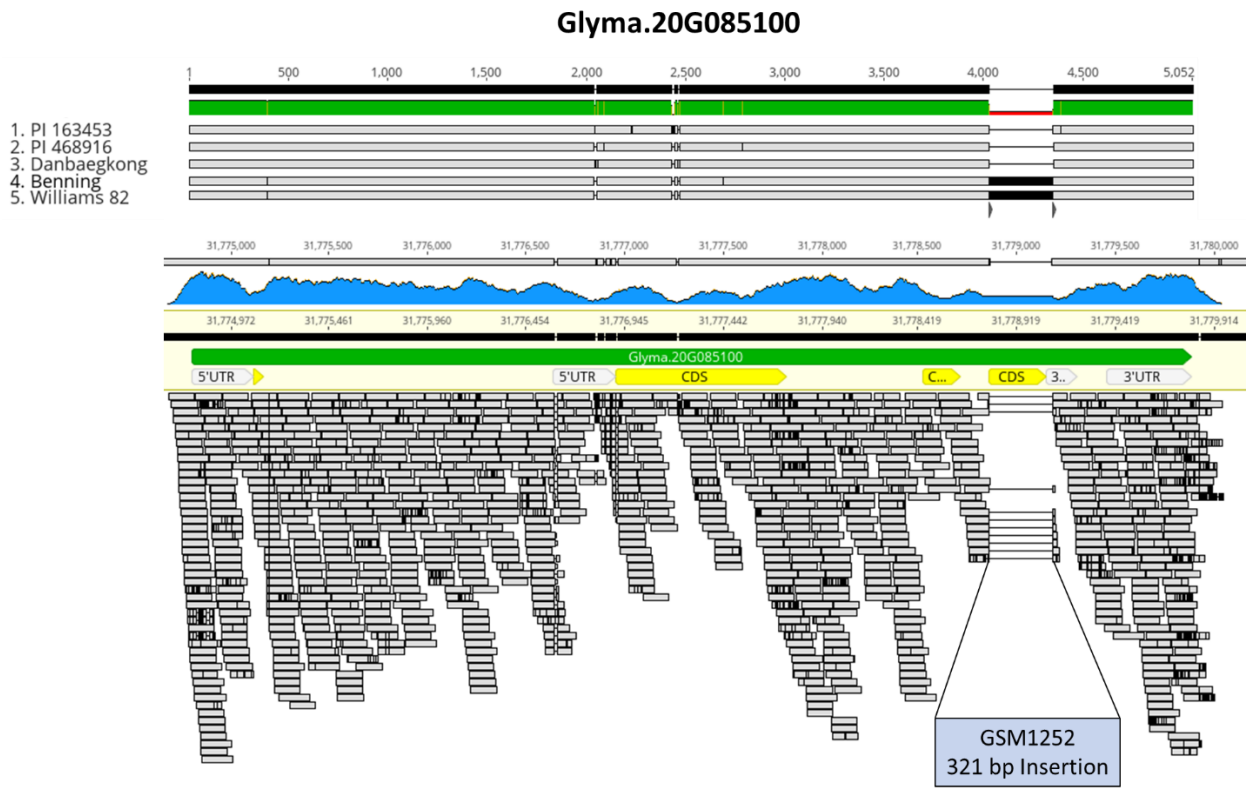

Figure 2.S5. Danbaekkong sequencing reads aligned to Williams 82.a2.v1 in the gene *Glyma.20g085100*. Sequence comparison of PI 163453, PI 468916, Danbaekkong, Benning, and Williams 82. The location of the TaqMan marker GSM1252 is indicated.

CHAPTER 3

MINING UNTAPPED GERMPLASM FOR GENETIC IMPROVEMENT OF PROTEIN QUANTITY AND QUALITY IN SOYBEAN [1]

1 Renan Souza, Blair Buckley, M. A. Rouf Mian, and Zenglu Li. To be submitted to *Plant Breeding*.

Abstract

Soybean meal is a major protein source for animal feeds, but its amino acid profile is not balanced, having low concentrations of cysteine (Cys) and methionine (Met). In this research, an exotic germplasm (PI 399000) was crossed with 'Woodruff' to develop a recombinant inbred line (RIL) population for mapping quantitative trait loci for seed composition. The population was grown in six environments and protein, oil, Cys, and Met concentrations were determined with near-infrared spectroscopy. RILs were genotyped with the SoySNP6K BeadChip and 1865 SNPs were used for QTL analysis. Entry mean-based heritability was 0.93, 0.92, and 0.82 for protein, oil and Cys + Met, respectively and a negative correlation between protein and Cys + Met was observed (-0.48). QTL analysis identified three loci on chromosomes (Chrs) 6, 15, and 17 in at least five environments for protein, six QTLs on Chrs 4, 6, 10, 14, 17, 19 in at least two environments for oil, three QTLs on Chrs 3, 6, and 10 for Cys and Met in at least two environments and two QTLs for seed size on Chrs 17 and 20 size in all environments. Stacking protein and Cys + Met QTLs in this population can increase both traits simultaneously and 13 breeding lines were identified with improved seed composition. The markers linked to the QTLs for high protein and elevated amino acid concentration can be used to develop soybean cultivars with improved soybean meal.

Key Words: Soybean meal, Seed composition, QTL mapping, Cysteine, Methionine

Introduction

Soybean has become a crop of importance on a global scale. Soybean meal is the main source of protein for animal feeds, while oil is a key component for human food, cosmetics, and biofuels. In 2021/2022 crop season, 28.5% of vegetable oil and 70% of protein meal used worldwide were derived from soybeans (USDA, 2023). Over the past 50 years, world soybean

production has increased approximately 10-fold, from 37.9 M tonnes in 1967 to 372 M tonnes in 2021 (FAO, 2023). The improvement in the standard of living in developing countries has resulted in an increase in the demand for animal derived products, consequently increasing the need of soybean protein for animal feeds (Guo et al., 2022). Another important factor is the increase in demand for biofuels and consequent growth in the production of vegetable oil (Qiu & Chang, 2010).

Soybean has a higher protein content than most cultivated plant species, but it is necessary to consider that quantity only is not important, but the value of the protein is in the quality. Quality of protein depends on a balanced distribution of essential and non-essential amino acids and in case of soybean, the content of sulfur amino acids is relatively low (Krishnan & Jez, 2018). Met, Cys, homocysteine, and taurine are four-common sulfur-containing amino acids, but only Met and Cys are proteogenic. Met and Cys are important in the growth and development of animals. Met is the amino acid responsible for the initiation of protein synthesis, while Cys is crucial for the formation of disulfide bonds (Brosnan & Brosnan, 2006). Content of these two amino acids in soybean is usually 2.6 g per 100 g of protein, which is lower than the recommended 3.5 g per 100 g of protein intake for monogastric animals (George & de Lumen, 1991; Shewry, 2000).

Increasing the levels of Cys and Met in soybean meal is important because these amino acids are essential for monogastric animals. In addition to Met and Cys, seven other amino acids are essential for poultry and swine (Lysine, Threonine, Tryptophan, Isoleucine, Arginine and Valine) (Baker, 2003; Boisen, 2003). Most of these amino acids are provided in soybean meal in sufficient amounts but Met and Cys are found in insufficient levels in the protein (Pfarr et al., 2018). To complement the low concentration of the sulfur-containing amino acids in soybean

meal, producers often add synthetic Met and Cys to meet the dietary requirements for optimal animal development, which results in an increase in production costs (Imsande, 2001; Krishnan, 2005). Development and use of new cultivars with improved amino acid content is important because it can increase the overall quality of meal and reduce the costs of livestock production (Durham, 2003).

Relationship between amino acid profile and overall seed protein content has been the focus of studies aiming to understand the effects of the increase in protein concentration. Fallen et al. (2013) found positive correlation between Met and Cys ($r = 0.76$) and these two amino acids also had a positive association with all other amino acids except for Lysine. Panthee et al. (2006a) reported that these two sulfur-containing amino acids were also positively associated with most of the other amino acids, except for valine. These results indicated that genetic improvements of Cys and Met in soybean can be achieved without drawbacks in the concentration of other essential amino acids.

Before a breeding strategy is adopted, it is necessary to understand the genetics underlying these traits of interests. In case of quantitative traits such as the amino acid content, QTL studies are essential to guide the development of breeding populations and selection process. Even though soybean protein quality is an important topic, genomic regions associated with individual amino acid concentration are limited in the literature (Panthee et al., 2006a; Panthee et al., 2006b; Fallen et al., 2013; Vaughn et al., 2014; Warrington et al., 2015; Lee et al., 2019) and the confirmed QTLs reported in the literature were only found in the reports by Panthee et al. (2006a and 2006b).

Panthee et al. (2006a) detected QTLs associated with Cys on chromosomes (Chrs) 1, 13, and 18, Met on Chrs 13, 18, and 7, and Met + Cys on Chrs 13 and 7. Another study by Panthee et

al. (2006b) identified loci associated with lysine concentration on Chrs 1, 15, and 18 and Threonine on Chrs 5, 2, 9, and 19. Fallen et al. (2013) reported three loci on Chr 13 associated with amino acid concentration. Two of these QTLs are very close to previously reported loci associated with protein content, which could be an indication that in some cases, QTLs for protein content can be associated with protein quality in soybean (Brummer et al., 1997; Reinprecht et al., 2006).

Warrington et al. (2015) identified a protein and a Cys + Met QTL in the same region on Chr 20 in the Benning × Danbaekkong population and, the favorable allele for the protein QTL was inherited from Danbaekkong, while the favorable allele for Cys + Met was contributed by Benning. This indicates a pleiotropic effect of this QTL which could limit its use to improve protein quantity and quality simultaneously. In a genome-wide association study conducted by Lee et al. (2019), 877 *G. max* in MGs I to IV were analyzed. The authors identified eight loci associated with essential amino acids, including Met on Chr 1, 15, and 18 and Cys on Chr 3, but the effects of these QTL were not stable across different environments. These studies demonstrated the challenge of combining quantity and quality of protein in breeding.

The first successful development of a soybean germplasm with improved amino acid composition was the outcome of the genetic mapping studies performed by Panthee et al. (2006a). One of the F_6-derived lines (TN04-5321) from the population was released as a new soybean germplasm with 43.1 % protein and 3.3% of Met and Cys (Panthee & Pantalone, 2006). Development of cultivars such as TN04-5321 is an important strategy to improve the nutritional quality of soybean because it can considerably increase the concentration of seed components. One of the limitations of breeding soybean lines with improved Cys + Met in soybeans is the narrow genetic variability for this trait (Patil et al., 2017). However, the genetic variation present

in germplasm collections can be explored to provide gradual improvements in these traits (Clarke & Wiseman, 2000).

In addition to the seed composition, another important trait for soybean breeding is the seed size. This trait along with the number of seeds determines the seed yield and is a determining factor for food grade classification of soybeans, which can be used as sprouts, edamame, or natto (Kato et al., 2014). Seed size is significantly correlated with weight, and this is routinely used as an indirect measurement in breeding programs. Seed weight can vary from 5.6 g to 34.8 g per 100 seeds in germplasm collections (Zhang et al., 2016; Zhao et al., 2019) and several QTLs have been reported in SoyBase (soybase.org) (Grant et al., 2009). The identification of QTLs for this trait in conjunction with the seed composition can provide valuable information for development of new soybean cultivars for specific uses.

Efforts to improve seed composition of soybean until now have been focused on increasing amount of protein or oil. Now, the importance of improving specific components such as essential amino acids is increasingly evident to develop higher quality soybean meal. In this context, present research aimed to 1) identify genomic regions controlling protein, oil, Met, Cys, and seed size and 2) develop molecular markers associated with the QTLs for marker-assisted selection.

Material and Methods

Plant materials and population development

A recombinant inbred line (RIL) population derived from Woodruff × PI 399000 (n=209) was used for mapping the seed composition in this study. Woodruff is a conventional maturity group (MG) VIII cultivar developed at the University of Georgia (Boerma et al., 2012). PI 399000 (MG V) is an accession originating from South Korea that has a high content of protein

as well as high Met and Cys content (Table 3.1). Hybridizations between two genotypes were performed in Athens, GA in 2016 and the F_1 generation was grown in the Illinois Crop Improvement Association (ICIA) nursery in Puerto Rico. F_2 seeds were planted in Athens during the summer of 2017. During the winter of 2017-2018, two cycles of a single seed descent method were used to advance the population to F_4 generation at the ICIA nursery in Puerto Rico. In 2018, the F_5 seeds were planted in Athens and plants were individually harvested and threshed to generate the RILs. The harvested $F_{5:6}$ seeds were planted in single rows (1.8 m x 0.8 m) in 2019. In 2020 and 2021, 209 RILs along with the parents were grown in three locations (Athens, GA, Bossier City, LA, and Hookerton, NC) with two replications in each location and a planting density of 27 seeds meter^{-1}. Plots dimension were 1.8 m x 0.8 m in Athens and 3.0 m x 1.0 m in Hookerton and Bossier City.

Trait evaluation

The content of protein (Pro), oil (Oil), Cys and Met was determined using near-infrared spectroscopy (DA 7250 Analyzer, PerkinElmer Inc., Waltham, MA, USA). The instrument was calibrated by the manufacturer using thousands of soybean samples with known seed composition across growing regions in the United States. The analysis was performed on a sample of approximately 200 seeds from each plot and seed composition was reported as percentage of seed on a dry matter basis. Amino acid contents were normalized to protein content for downstream analysis by dividing the measured content of each amino acid by the total measured protein content for each sample. Concentrations of Cys and Met were summed to determine the total amount of sulfur-containing amino acids (Cys+Met). Seed size was also assessed as the weight (g) per 100-seeds from each plot harvested in all environments.

Statistical analysis

Phenotypic data was analyzed in RStudio with the R version 4.1.0 (R Core Team, 2021). Packages lme4 (Bates et al., 2015), lmertest (Kuznetsova et al., 2017), and VCA (Schuetzenmeister & Dufey, 2020) were used for the phenotypic data analysis. The package ggplot2 was used for visualization (Wickham, 2016). Best linear unbiased predictions (BLUPs) were calculated from single and multi-environments (years + location) and the values for each RIL was obtained with the following statistical model (Henderson, 1975):

$$y = X_1 m + X_2 l + X_3 r + Z_1 g + Z_2 i + e$$

y is the response vector, m is the vector of the overall mean, l is the vector of the environment (location + year) (fixed), r is the vector of the replication within environment (fixed), g is the vector of the RILs (random), i is the vector of the interaction RIL x Environment effect (random), and e is the vector of residuals (random). X and Z represent incidence matrices of the m, l, r, g, i effects, respectively.

The variance components were determined in a model considering all factors as random effects using the restricted maximum likelihood method (REML) (Patterson & Thompson, 1971). These components were used to determine entry-mean based heritability estimates according to the following equation (Nyquist & Baker, 2008):

$$h^2 = \sigma_g^2 / (\sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma_\varepsilon^2}{re})$$

h^2 is the heritability, σ_g^2 genotypic variance, σ_{ge}^2 is the G×E variance, σ_ε^2 error variance, r is number of replicates.

DNA extraction and genotyping

A sample of 20 seeds from each RIL was ground in a Perten 3310 laboratory Mill (PerkinElmer Inc., Waltham, MA, USA) and ~100 mg of seed tissue was used for DNA extraction using the Edwards et al. (1991) protocol. A total of 209 RILs along with the parents were genotyped using the Illumina SoySNP6K iSelect BeadChip (Illumina, San Diego, CA, USA) (Song et al., 2013). A quality control step was performed to remove monomorphic markers and markers with inconsistent genotype calls. SNPs with significant segregation distortion, and call rate <90% were removed, resulting in 1865 SNPs.

Linkage mapping and QTL analysis

Map construction and linkage analysis were performed with the R package R/qtl (Broman et al., 2003). Recombination frequencies were initially estimated with the function est.rf() and a map was constructed with the function est.map() using Kosambi's mapping function and the significance level of 0.001. Linkage groups were assigned to a chromosome number based on soybean reference genome Williams82.a2.v1 (soybase.org). QTL analysis was performed using composite interval mapping function cim() considering a window of 10 cM and QTL were confirmed when LOD values were higher than a LOD significance threshold estimated with 1000 permutations for each trait. QTL confidence interval was determined with the function bayesint() from the package R/qtl.

Results

The content of protein, oil and amino acids was measured using near-infrared spectroscopy (NIR) and seed size was determined from 209 RILs derived from the population Woodruff × PI 399000 evaluated at three locations in 2020 and 2021. Analysis of variance indicated that genotype, environment, and G × E interaction had significant effects for all traits

evaluated. Genotype was the most important source of variation for protein, oil, and seed size, but for Cys and Met, environment and genotype had similar effects in the variation presented in the population (Table 3.2).

Entry mean-based heritability ranged from 0.78 for Cys to 0.95 for seed size and all traits exhibited a relatively high estimation, indicating a strong effect of the genetic component. The coefficient of variation was low for all traits, ranging from 1.82% for Met to 5.35% for seed size indicating good experimental precision in the trials (Table 3.2). The values from each plot for protein ranged from 37.0 to 47.0 % across the six environments, while the values for oil ranged from 15.9 to 23.0 %. The values for Cys + Met ranged from 2.54 to 2.99% and for seed size, the range was 9.3 to 23.8 g/100-seed (Table 3.3). The population exhibited a broader variation in comparison to the parental mean and followed a normal distribution (Figure 1). The broad variation in the progeny with several RILs exceeding the parental values for the traits indicates the presence of transgressive segregation. The recombination between parental genotypes that possess QTLs with antagonistic effects can create new combinations of favorable alleles that will increase the value for each trait in the progeny.

The collection of phenotypic data in six environments gives an opportunity to explore the effect of the environment on the traits evaluated. The environment with the highest mean protein was LSU20 (43.7%) and the lowest was ATH21 (41.8%). For oil content, the environment with the highest mean was NC21 (19.9%) and the lowest was LSU21 (19.4%). In the case of Cys+ Met, the best environment was ATH20 (2.8%) and environment with the lowest mean was LSU21 (2.7%). For seed size, LSU20 produced the largest seeds (16.6 g per 100-seed) and NC21 had the smallest seeds (13.8 g per 100-seed) (Table 3.S1). Across all environments, the protein content was negatively correlated with oil (-0.5**) and Cys + Met (-0.48**) and positively

correlated with seed size (0.35**). Oil was positively correlated with Cys + Met (0.26**) and negatively correlated with seed size (-0.16**). Cys + Met had a low correlation with seed size (-0.09**) (Table 3.4).

QTLs for seed composition traits

After removing markers with segregation distortion and redundancy, the map construction was performed with 1865 SNPs. In total, 20 linkage groups were formed with a total length of 2543.5 cM (Table 3.S2). In total, 22 QTLs were found across nine chromosomes (Chrs 3, 4, 6, 10, 14, 15, 17, 19, and 20) for all traits. Of those, Chrs 6 and 10 contained the highest number of QTLs (qPro-6, qOil-6,qMet-6,qCys-6 and qOil-10, qMet-10,qCys-10) (Table 3.5, Figure 3.S1).

Three QTLs (qPro-6, qPro-15 and qPro-17) were identified for protein. The QTL qPro-6 was identified across all environments, while the qPro-15 and qPro-17 were detected in 5 out of 6 environments except for ATH20 and NC20, respectively. These QTLs together explained 52.6% of the phenotypic variation for protein, with the qPro-6 alone accounting for 20.3%. The favorable alleles for elevated protein content on qPro-6 and qPro-15 were from PI 399000, while the favorable allele of qPro-17 is from Woodruff (Figure 3.2, Table 3.5). Interaction between the top two QTLs for protein content (qPro-6 and qPro-15) was observed. When a genotype carries the alleles of PI 399000 at both loci, the protein content increases more than the additive value of the two QTLs, going from 42.0 to 43.7% (Table 3.S3). In addition to the three QTLs (Chrs 6, 15 and 17), four QTLs with minor effects (LOD > 2.5, $R^2 < 10$) were also identified (Chrs 2, 9, 10, and 12) with favorable alleles on Chr 9, 10, and 12 from PI 399000 and on Chr 2 from Woodruff. Together, these seven loci explained a total of 70% of phenotypic variance for protein content (Figure 3.S3).

In total, six QTLs (qOil-4, qOil-6, qOil-10, qOil-14, qOil-17 and qOil-19) were identified for oil, with LOD scores ranging from 4.2 to 10.8. The qOil-6 overlaps with the protein QTL qPro-6, indicating a pleiotropic effect of this locus (Figure 3.S1). In total, these six QTLs explained 72.5 % of the phenotypic variation, with the qOil-14 accounted for 20.4% of phenotypic variation and contributed 0.31% of the oil content (Table 3.5). The qOil-10, 14, 17, and 19 had favorable alleles contributed by PI 399000. The qOil-14 and qOil-17 were identified in all six environments and the other QTLs were identified in two to four environments.

QTL analysis was performed for Cys and Met separately and in combination as a single trait (Cys+Met). Four QTLs were identified for Met (qMet-3, 6, 10 and 15) (Figure 3.2, Table 3.5) and the combined variation explained by these QTLs was 54.2%. The qMet-3 and qMet-10 explained 11.3 and 15.2% of phenotypic variation, respectively and the favorable alleles at both loci were from PI 399000. The favorable alleles at qMet-6 and qMet-15 were from Woodruff and these loci explained 16.0 and 11.8% of the variation, respectively. The QTLs qMet-6, 10 and 3 were identified in four, three and two environments, respectively. The qMet-15 was identified only in the combined analysis of all environments, not being detected in any individual environment.

For Cys, three QTLs were identified (qCys-6, 10, 15). The QTLs qCys-10 and qCys-15 were located in the same region as the qMet-10 and qMet-15, respectively. In the combined QTL analysis of cysteine + methionine, loci were detected on Chrs 3, 6, 10 and 15. The qCM-3 QTL was in the same interval as qMet-3 and qCM-6, qCM-10, qCM-15 were in the same location as qCys-6, qCys-10 and qCys-15, respectively. In addition to the four QTLs identified for cysteine and methionine, another two loci with LOD > 2.5 were identified on Chr 14 and 20 and in total, these QTLs explained 65.4% of the variance.

The QTL analysis for seed size indicated the presence of two QTLs (Figure 3.2). The QTL qSS-17 explained 21.9% of the variance and the allele from Woodruff increased the seed size by 0.81g/100-seed. The other QTL for seed size (qSS-20) was also contributed by Woodruff and explained 14.8% of the variance, increasing the seed size by 0.68 g/100-seed. In addition to the two QTLs for seed size, 10 other QTLs with minor effect were detected on Chrs 3, 4, 5, 7, 9, 11, 13, 15, 16, and 19 and the cumulative effect of these loci explained 88.0% of the phenotypic variance. (Figure 3.S3).

Discussion

Genetic variation of seed composition

The protein content of soybean in the United States has been decreasing in the last years because more focus was given on yield increases than seed composition (Naeve & Miller-Garvin, 2019). To improve the protein content of soybeans, plant breeders began developing new germplasm with increased protein content (Wilcox & Cavins, 1995; Burton & Wilson, 1998; Mian et al., 2008). However, it was observed that the selection for increased crude protein content has been associated with a decrease in the content of sulfur-containing amino acids. This is because the increase was mostly in non-essential amino acids, such as arginine and glutamic acid (Pfarr et al., 2018). Therefore, a more effective strategy needs to combine the selection for high protein with high cysteine and methionine content simultaneously.

In order to provide gains in quantity and quality of protein, genetic diversity of germplasm can be explored for development of new breeding lines. Analyzing the variation for protein content and sulfur amino acids in the USDA Soybean Germplasm Collection, 254 *G. max* accessions in MGs V to X have protein above 44% and Cys + Met above 3%. Of these 254 accessions, 116 accessions came from North and South Korean (Figure 3.S4). It has been

observed that soybeans from South Korea usually have a higher concentration of seed protein than those from United States (Vaughn et al., 2014; Bandillo et al., 2015; Patil et al., 2017). This is likely a result of historical breeding efforts that focused on the improvement of protein for the production of soyfoods, such as tofu (Lee et al., 2015). In this context, accessions from South Korea are an important source of germplasm to improve the protein quantity and quality of soybean because they carry favorable alleles for these traits. PI 399000 is a Korean accession used in the present research that combines the high concentrations of protein with Met and Cys and was chosen to develop a population to study the genetic control of these traits and assess the possibility of having high protein and high sulfur-containing amino acids content simultaneously.

Effect of environments on seed composition

The performance of the population among the trials for all traits evaluated and this revealed how different environmental conditions shape the final phenotypes. The environments with the highest mean protein content were LSU20 and LSU21, while the lowest values of oil were also identified in Louisiana. This indicates that the environmental factors that are favorable for high protein production have a detrimental effect on oil. Analyzing weather information for the environments, LSU20 and LSU21 had the highest average temperatures and the lowest precipitation during the growing season in comparison with the other locations. Dornbos & Mullen (1992) indicated that higher temperatures during the seed filling stage leads to higher protein content. Similarly, Patil et al. (2017) indicated that higher temperatures positively affect the protein content but lead to a reduction in oil. The sulfur amino acid content was stable overall indicating that this trait is not largely affected by the environment. Different results were obtained by Mourtzinis et al. (2017) where the environments with higher temperature during the R5 to R8 stage had a reduction in the concentration of essential amino acids.

The seed size had variation across the tested locations, with the highest values found in Louisiana (LSU20) and the lowest values in the trials in North Carolina. Analyzing the weather data, it was noted that NC20 and NC21 have the lowest average temperature during the growing season and the highest precipitation, which indicates that these environmental factors might play a role in determining seed size. However, Choi et al. (2016) did not observe significant differences in seed size when growing soybean cultivars at different temperature treatments.

Despite the fact of the environmental effects, the majority of the identified QTLs with large effects showed stability. The protein QTLs qPro-6, 15 and 17 were identified in at least 5 of the 6 environments and the top two QTLs for oil were identified in all environments. The QTL for cysteine + methionine qCM-6 was identified in 5 environments and the two QTLs for seed size (qSS-17 and qSS-20) were detected in all environments. These results demonstrate that use of these highly stable QTLs for marker assisted selection may facilitate selection across different environmental conditions.

Heritability and correlation among traits

The heritability estimates for the traits were high, with protein at 0.93, oil at 0.92, Cys + Met at 0.82 and seed size at 0.95. The value for protein is similar to the estimation of Warrington et al. (2015), where protein also had 0.93, but the estimation for Cys + Met in the present study was higher than the value that was reported (0.45-0.59). The value calculated here was also higher than the estimate by Fallen et al. (2013), who reported the heritability for Cys was 0.63 and for Met was 0.67. For seed size, studies have shown that the heritability can reach 0.98, a value similar to the estimate from this study, suggesting that genetic components are very important in controlling the trait (Zhang et al., 2016; Yan et al., 2017).

The analysis of the phenotypic data showed a negative correlation between protein and other composition traits, including oil (-0.5**), Cys (-0.47**) and Met (-0.37**). Arnold et al. (2021) indicated a negative correlation between protein and oil (-0.54** to -0.72**) and a negative correlation between protein and Cys (ranging from -0.12 to -0.16**) and Met (ranging from -0.31** to -0.43**). Similarly, Warrington et al. (2015) found a negative correlation between protein content and Met (-0.19*) and Cys (-0.16*) and a high positive correlation between Met and Cys (0.57**). Panthee et al. (2006a) found a negative correlation between protein and Cys (-0.48**), however, the authors did not detect significant correlation between protein and Met. Seed size was positively correlated with protein content (0.35**) and negatively correlated with all other seed composition traits evaluated. A positive correlation between seed size and protein content was also indicated by Alt et al. (2002) and Panthee et al. (2005) and this can be an indication that the selection of genotypes with larger seed size might lead to increased protein content.

Comparison with previous reported QTLs of seed composition

In this research, QTLs for protein were identified on Chrs 6, 15, and 17. Arnold et. al. (2021) identified a protein QTL on Chr 6 between markers Gm06_13990118 and Gm06_13732043 which is 4 Mb downstream the location of the qPro-6 QTL found in this study (Gm06_9081030_C_A - Gm06_9668798_T_C). Due to the distance between the QTLs, it can be determined that they are different QTLs. The qPro-15 identified in this study is located 15.3 Mb downstream a QTL identified by Warrington et al. (2015) and 18.1 Mb downstream a well-known QTL on Chr 15 (Kim et al., 2016), suggesting that they are likely different QTLs. Warrington et al. (2015) reported a QTL on chromosome 17 located 31.1 Mb downstream the locus identified in the present research, which indicated that the qPro-17 detected in this study

might be novel. Significant interaction between the top two QTLs (qPro-6 and qPro-15) was observed for protein content. Qi et al. (2016) and Zhang et al. (2018) also reported epistatic effects for protein QTLs in soybean, indicating that QTL × QTL interaction is important in determination of the trait. To identify candidate genes underlying the protein QTLs, a gene search was performed in a 100 kb window centered on the most significant marker at each QTL with largest effect, seven, one and seven candidate genes were identified for the qPro-6, qPro-15, and qPro-17 QTLs, respectively (Supplementary Table S4).

Six QTLs associated with the oil content were identified in the present research. Two of these QTLs, qOil-14 and qOIL-17 were detected in all six environments. The QTL qOil-14 was located between 45.5 and 45.9 Mb and two genes that were previously reported in this region are associated with oil content in soybeans (Fang et al., 2017). One of these genes is *Glyma.14G194300* that is a Fatty acid desaturase 3 located at 45.93 Mb and other one is *Glyma.14G193800* which is a Phospholipase C 2 located at 45.86 Mb. The QTL qOil-17 was located between 7.83 and 8.00 Mb, and there are 27 genes in this interval, but none of them have been reported in the literature in association associated with oil. However, at the 8.91 Mb position, there is a diglyceride acyltransferase gene DGAT (*Glyma.17g112800*) that was associated with increase in oil content in soybean (Roesler et al., 2016).

Four QTLs were identified (Chrs 3, 6, 10, and 15) for sulfur-containing amino acids. The qCys-6 QTL was identified 0.46 Mb downstream from qMet-6 and qCys-10 and qCys-15 were identified in the same interval as the QTLs qMet-10 and qMet-15, respectively. This can be an indication that the loci that control these two amino acids might be the same. When analyzing the two traits combined (CM), the QTL regions were the same as those when the traits were

analyzed separately on Chrs 6, 10 and 15. This suggests that the control of these two amino acids might be the same.

The QTL for Cys+Met qCM-6 is located at the 10 Mb position, which is approximately 7 Mb upstream an QTL found in the same chromosome by Arnold et al. (2021) and 7.51 Mb downstream a QTL identified by Warrington et al., (2015) and Singer et al. (2022), which might indicated that it is a different QTL. In the present study, a QTL for Cys + Met was also found on Chr 3 between 5 and 13 Mb. Singer et al. (2022) identified a locus on Chr 3 between 33.8 and 41.2 Mb and Lee et al. (2019) identified a QTL at 42.7 Mb for Cys in a GWAS study, indicating that the region identified here is a novel locus. The QTL qCM-15 was identified between 22 and 47 Mb and this region overlaps with a QTL identified by Lee et al. (2019) located between 38 and 39 Mb.

The QTLs for Met + Cys qCM10 identified here overlap with the loci reported by Arnold et al. (2021). Since a Met + Cys QTL has been repeatedly identified on Chr 10 between 3 and 4 Mb, it is likely that there is a gene associated with the control of sulfur-containing amino acids.Boehm et al. (2018) mapped a QTL controlling the 11S protein subunit on Chr 10 between 2.7 and 3.3 Mb that overlaps with the qCM10 found here. In this region there is one glycinin gene *Gy4* that encodes the 11S globulin subunit (Diers et al. 1993). The 11S storage proteins has approximately four times more sulfur-containing amino acids per gram of protein than the 7S subunit. In this context, it is likely that the qCM10 is controlled by the gene *Gy4.*

The QTLs on Chr 6, 10 and 15 identified for cysteine were located in the same region as the QTL identified formethionine. It is likely that they are the same QTL since these two amino acids share the same biosynthetic pathway. Plants assimilate sulfate from the soil and convert it to sulfide, which is then added to O-Acetylserine and gives rise to Cys, which is the precursor of

Met (Ravanel et al., 1998). In total, 16 genes were identified in the region of the QTLs detected in more than 5 environments (Chr 6, 10, and 15), with nine genes on Chr 6, six genes on Chr 10 and one on Chr 15.

The qPro-6 and qMet-6 were detected in the same location. Overlap between protein QTLs and specific amino acids is expected, since the amino acids are the building blocks for protein (Panthee et al., 2006a). The favorable allele for qPro-6 QTL is from PI 399000, however, the favorable alleles for qMet-6 at the locus is from Woodruff. The fact that the favorable alleles for protein and sulfur amino acids come from different parents at the same locus indicates a pleiotropic effect of the QTL that might complicate the effort to combine high protein and high sulfur amino acid. The pleiotropic effect of these QTLs on protein and sulfur-containing amino acids is likely due to a dilution effect. In this effect, the higher protein content in soybean seeds in high protein genotypes is due to increased content of non-essential amino acids, such as arginine and glutamic acid. This leads to an overall a reduction of the share of essential amino acids in the protein (Zarkadas et al., 2007; Serretti et al., 1994; Pfarr et al., 2018). In a similar observation, in the present research, the qProt-15 and qMet-15 which are located in the same region, have favorable alleles coming from different parents.

The allele from PI 399000 at qProt-15 increases the total protein, but the allele from Woodruff at qMet-15 increases the methionine content. To reduce the negative pleiotropic effects, a good option to increase both protein and amino acid might be to stack the QTLs in different regions, such as the qPro-17 with qCM-3 and qCM-10. In fact, it was found six breeding lines in this population with this specific combination of alleles that present a protein content higher than 43% and cysteine + methionine higher than 2.8 % (Table 3.S5).

QTLs for seed size were found on Chrs 17 and 20. The QTL qSS-17 was found between 7.19 and 7.24 Mb that is approximately 1.67 Mb upstream a well-known gene controlling seed size in soybean, the *GmKIX8-1* (Nguyen et al., 2021). The QTL qSS-20 was found between 41.47 and 42.24 Mb, which is near a region harboring a seed size QTL previously reported (Qi et al., 2020).The fact that seed size QTLs on Chrs 17 and 20 have been reported and were also detected in the present research in all environments studied, indicate that these loci are very stable in the control of the trait and can be targeted for breeding selection.

Challenges on genetic improvement of sulfur-containing amino acids

To quantify the total variance explained by all possible QTLs underlying the traits studied, a list of all possible QTLs with a LOD higher than 2.5 was obtained from the mapping results. The most significant marker at each location was fitted in a linear model to estimate the share of the phenotypic variance that can be attributed to additive marker effects. This was to understand how many genomic regions are associated with each trait evaluated. For protein, seven regions associated with protein were identified. These seven QTLs explained a total of 70% of variance. For oil, eight regions explained a total of 71.3% of the variance and for Cys + Met, six putative regions explained 65.4% of the variance in the population. Based on the results, it is possible to deploy a MAS for seed composition traits since most of the phenotypic variance can be explained by relatively few loci in the genome. However, in the case of Cys+ Met, each QTL presented a small effect which represent a challenge to accelerate breeding of these amino acids and reach the optimal content of 3.5% in the protein (George & de Lumen, 1991). A strategy in the case of several QTLs with minor effect is to deploy whole genome prediction to capture the variation of all loci and accumulate favorable alleles for sulfur-containing amino acids (Miller et al. 2023).

Breeding with diverse germplasm for seed composition

In addition to studying the genetic control of important traits in soybeans, this research also aimed to introduce novel germplasm in an effort to broaden the genetic base of soybean. This an important effort because of the relatively low diversity of the North American soybean gene pool (Gizlice et al., 1994). The domestication process of soybean in Asia and the introduction in North America have caused several genetic bottlenecks and estimates indicate that approximately 80% of the rare alleles have been eliminated in modern soybean in comparison with wild ancestors (Hyten et al., 2006).

There have been efforts to incorporate exotic germplasm into elite lines to increase the relative genetic diversity (Carter et al., 2007), but in general, the incorporation of exotic materials is associated with reduction in agronomic performance and yield (Vello et al., 1984). However, it is important to consider that the new germplasm brings favorable traits for seed composition and stress resistance that are becoming more relevant as soybean uses evolve and the climate continues to change.

Previous research has been done in efforts to combine high protein with increased concentration of amino acids, but the QTLs found for protein had an effect in decreasing the essential amino acids (Warrington et al, 2015). In the present research it was observed that different combinations of protein QTLs and Cys + Met QTLs can provide gains in both components. It was possible to identify 13 breeding lines from this population with protein content higher than 43% and sulfur amino acid content higher than 2.8% with at least 2 protein QTLs and 2 Cys + Met QTLs in coupling phase. These results provide resources for improving protein quantity and quality of soybean germplasm in the United States.

Conclusions

Soybean meal is the largest source of animal protein feed because of the high protein concentration, availability, and low cost. However, improvements are needed in the amino acid composition of the protein. In the present research, 22 QTLs were identified, several of them being novel that were associated with protein, Cys and Met. The Pleiotropic effect between protein and Cys+Met QTLs on Chr 6 and 15 remain a challenge to stack both traits but the combination of favorable alleles from different QTLs via MAS has the potential to increase the protein without decreasing concentration of Cys + Met. The markers linked to the QTLs for high protein and elevated amino acid concentration can be used to study the genetic control of the traits and to accelerate breeding of soybean cultivars with improved soybean meal. Use of exotic germplasm is an important strategy to further improve soybean seed composition and to increase the genetic diversity of soybeans in North America.

Acknowledgments

We thank Tatyana Nienow, Nicole Bachleda, Dale Wood, Brice Wilson, and Brian Little at the University of Georgia for the technical support.

References

Alt, B.J., Fehr, W.R., & Welke, G.A. (2002). Selection for large seed and high protein in two- and three-parent soybean populations. *Crop Science* 42(6): 1876–1881. https://doi.org/10.2135/cropsci2002.1876.

Arnold, B., Menke, E., Mian, M.A.R., Song, Q., Buckley, B., & Li, Z. (2021). Mining QTLs for elevated protein and other major seed composition traits from diverse soybean germplasm. *Molecular Breeding* 41(8): 1–18. https://doi.org/10.1007/s11032-021-01242-z.

Baker, D. (2003). Ideal amino acid patterns for broiler chicks. In: D'Mello, J., editor, Amino

Acids in Animal Nutrition. 2nd ed. CABI, Cambridge, MA. p. 223–235.

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., & Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *The Plant Genome* 8(3): 1–13. https://doi.org/10.3835/plantgenome2015.04.0024.

Bates, D., Mächler, M., Bolker, B.M., & Walker, S.C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Boehm, J.D., Nguyen, V., Tashiro, R., Anderson, D., Chun, S., Wu, X., Woodrow, L., Yu, K., Cui, Y., Li, Z. (2018). Genetic mapping and validation of the loci controlling 7S α ′ and 11S A - type storage protein subunits in soybean [*Glycine max* (L .) Merr .]. *Theoretical and Applied Genetics*. 131: 659–671. doi: 10.1007/s00122-017-3027-9.Boerma, H.R., Hussey, R.S., Phillips, D. V, & Wood, E.D. (2012). Soybean Variety G00-3209. US Patent 8304616 B2.

Boisen, S. (2003). Ideal dietary amino acid profiles for pigs. In: D'Mello, J., editor, Amino Acids in Animal Nutrition. 2nd ed. CABI, Cambridge, MA. p. 157–168

Broman, K.W., Wu, H., Sen, Ś., & Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19(7): 889–890. https://doi.org/10.1093/bioinformatics/btg112.

Brosnan, J., & Brosnan, M. (2006). 5th Amino Acid Assessment Workshop. *The Journal of Nutrition* 136(6): 1636–1640.

Brummer, E.C., Graef, G.L., Orf, J., Wilcox, J.R., & Shoemaker, R.C. (1997). Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Science* 37(2): 370–378. https://doi.org/10.2135/cropsci1997.0011183X003700020011x.

Burton, J., & Wilson, T. (1998). Registration of ‘Prolina’soybean. *Crop Science* 39: 294–295.

Carter, T.E., Burton, J.W., Fountain, M.O., Rzewnicki, P.E., Villagarcia, M.R., & Bowman, D.T. (2007). Registration of ‘N7002’ Soybean. *Journal of Plant Registrations* 1(2): 93–94. https://doi.org/10.3198/jpr2006.12.0830crc.

Choi, D.H., Ban, H.Y., Seo, B.S., Lee, K.J., & Lee, B.W. (2016). Phenology and seed yield performance of determinate soybean cultivars grown at elevated temperatures in α temperate region. *PLoS ONE* 11(11): 1–18. https://doi.org/10.1371/journal.pone.0165977.

Clarke, E., & Wiseman, J. (2000). Developments in plant breeding for improved nutritional quality of soya beans I. Protein and amino acid content. *The Journal of Agricultural Science* 134(2): 111–124.

Diers, B.W., Beilinson, V., Nielsen, N.C., Shoemaker, R.C. (1993) Genetic mapping of the Gy4 and Gy5 glycinin genes in soybean and the analysis of variant Gy4. *Theoretical and Applied Genetics* 89:297–304

Dornbos, D.L., & Mullen, R.E. (1992). Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. *Journal of the American Oil Chemists Society* 69(3): 228–231. https://doi.org/10.1007/BF02635891.

Durham, D. (2003). The United Soybean Board’s better bean initiative: Building United States soybean competitiveness from the inside out. *AgBio Forum* 6(1): 23–26.

Fallen, B.D., Hatcher, C.N., Allen, F.L., Kopsel, D.A., Saxton, A.M., Chen, P., Kantartzi, S.K., Cregan, P.B., Hyten, D.L., & Pantalone, V.R. (2013). Soybean seed amino acid content QTL detected using the Universal Soy Linkage Panel 1.0 with 1,536 SNPs. *Journal of Plant Genome Sciences* 1(3): 68–79. https://doi.org/10.5147/jpgs.2013.0089.

Fang, C., Ma, Y., Wu, S., Liu, Z., & Wang, Z. (2017). Genome-wide association studies dissect

the genetic networks underlying agronomical traits in soybean. *Genome Biology* 18(1). https://doi.org/10.1186/s13059-017-1289-9.

FAO. (2023). FAOSTAT - Statistics Database. http://www.fao.org/faostat/en/#home (accessed 2 March 2023).

George, A.A., & de Lumen, B.O. (1991). A novel methionine-rich protein in soybean seed: Identification, amino acid composition, and N-terminal sequence. *Journal of Agricultural and Food Chemistry* 39: 224–227.

Gizlice, Z., Carter Jnr, T.E., & Burton, J.W. (1994). Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Science* 34(5): 1143–1151. https://doi.org/10.2135/cropsci1994.0011183X003400050001x.

Grant, D., Nelson, R.T., Cannon, S.B., & Shoemaker, R.C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 38: 843–846. https://doi.org/10.1093/nar/gkp798.

Guo, B., Sun, L., Ren, H., Sun, R., Wei, Z., Hong, H., Luan, X., Wang, J., Wang, X., Xu, D., Li, W., Guo, C., & Qiu, L. (2022). Soybean genetic resources contributing to sustainable protein production. *Theoretical and Applied Genetics* 135, 4095–4121. doi: 10.1007/s00122-022-04222-9.

Henderson, C.R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* 31(2): 423–447.

Hyten, D.L., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., & Cregan, P.B. (2006). Impact of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences of the United States of America* 103(45): 16666–16671. https://doi.org/10.1073/pnas.0604379103.

Imsande, J. (2001). Selection of Soybean Mutants with Increased Concentrations of Seed Methionine and Cysteine. *Crop Science* 41: 510–515. https://doi.org/10.2135/cropsci2001.412510x.

Kato, S., Sayama, T., Fujii, K., Yumoto, S., Kono, Y., Hwang, T.Y., Kikuchi, A., Takada, Y., Tanaka, Y., Shiraiwa, T., & Ishimoto, M. (2014). A major and stable QTL associated with seed weight in soybean across multiple environments and genetic backgrounds. *Theoretical and Applied Genetics* 127(6): 1365–1374. https://doi.org/10.1007/s00122-014-2304-0.

Krishnan, H.B. (2005). Engineering Soybean for Enhanced Sulfur Amino Acid Content. *Crop Science* 45: 454–461. https://doi.org/10.2135/cropsci2005.0454.

Krishnan, H.B., & Jez, J.M. (2018). Plant Science Review : The promise and limits for enhancing sulfur-containing amino acid content of soybean seed. *Plant Science* 272(1): 14–21. https://doi.org/10.1016/j.plantsci.2018.03.030.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software; 82(13), 1-26*. https://www.jstatsoft.org/v082/i13.

Lee, C., Choi, M., Kim, H., Yun, H., Lee, B., Chung, Y., Kim, R., & Choi, H. (2015). Soybean [Glycine max (L.) Merrill]: Importance as A Crop and Pedigree Reconstruction of Korean Varieties. *Plant Breeding and Biotechnology.* 3(3): 179–196. https://doi.org/10.1016/S0828-282X(08)70684-6.

Lee, S., Van, K., Sung, M., Nelson, R., LaMantia, J., McHale, L.K., & Mian, M.A.R. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theoretical and Applied Genetics* 132(6): 1639–1659. https://doi.org/10.1007/s00122-019-03304-5.

Mian, M.A.R., Cooper, R.L., & Dorrance, A.E. (2008). Registration of 'Prohio' Soybean. *Journal of Plant Registrations* 2(3): 208. https://doi.org/10.3198/jpr2007.09.0531crc.

Miller, M. J., Song, Q., & Li, Z. 2023. Genomic Selection of Soybean (*Glycine max*) for Genetic Improvement of Yield and Seed Composition in a Breeding Context. *Plant Genome*. https://doi.org: 10.1002/tpg2.20384

Mourtzinis, S., Gaspar, A.P., Naeve, S.L., & Conley, S.P. (2017). Planting date, maturity, and temperature effects on soybean seed yield and composition. *Agronomy Journal* 109(5): 2040–2049. https://doi.org/10.2134/agronj2017.05.0247.

Naeve, S., & Miller-Garvin, J. (2019). United States soybean quality - Annual Report. Dep. of Agronomy, University of Minnesota, St. Paul.

Nguyen, C.X., Paddock, K.J., Zhang, Z., & Stacey, M.G. (2021). GmKIX8-1 regulates organ size in soybean and is the causative gene for the major seed weight QTL qSw17-1. *New Phytologist* 229(2): 920–934. https://doi.org/10.1111/nph.16928.

Nyquist, W.E., & Baker, R.J. (2008). Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences* 10(3): 235–322.

Panthee, D.R., & Pantalone, V.R. (2006). Registration of Soybean Germplasm Lines TN03–350 and TN04–5321 with Improved Protein Concentration and Quality. *Crop Science* 46(5): 2328–2329. https://doi.org/10.2135/cropsci2005.11.0437.

Panthee, D.R., Pantalone, V.R., Sams, C.E., Saxton, A.M., West, D.R., Orf, J.H., & Killam, A.S. (2006a). Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. *Theoretical and Applied Genetics* 112(3): 546–553. https://doi.org/10.1007/s00122-005-0161-6.

Panthee, D.R., Pantalone, V.R., Saxton, A.M., West, D.R., & Sams, C.E. (2006b). Genomic

regions associated with amino acid composition in soybean. *Molecular Breeding* 17(1): 79–89. https://doi.org/10.1007/s11032-005-2519-5.

Panthee, D.R., Pantalone, V.R., West, D.R., Saxton, A.M., & Sams, C.E. (2005). Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. *Crop Science* 45(5): 2015–2022. https://doi.org/10.2135/cropsci2004.0720.

Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., Shannon, G.J., Carter, T.C., & Nguyen, H.T. (2017). Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. *Theoretical and Applied Genetics* 130(10): 1975–1991. https://doi.org/10.1007/s00122-017-2955-8.

Patterson, H.D., & Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika* 58(3): 545–554.

Pfarr, M.D., Kazula, M.J., Miller-Garvin, J.E., & Naeve, S.L. (2018). Amino acid balance is affected by protein concentration in soybean. *Crop Science* 58(5): 2050–2062. https://doi.org/10.2135/cropsci2017.11.0703.

Qi, Z., Pan, J., Han, X. (2016). Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. *Mol Breeding* 36(55).

Qi, Z., Song, J., Zhang, K., Liu, S., Tian, X., Wang, Y., Fang, Y., Li, X., Wang, J., Yang, C., Jiang, S., Sun, X., Tian, Z., Li, W., & Ning, H. (2020). Identification of QTNs Controlling 100-Seed Weight in Soybean Using Multilocus Genome-Wide Association Studies. *Frontiers in Genetics* 11: 1–12. https://doi.org/10.3389/fgene.2020.00689.

Qiu, L.J., & Chang, R.Z. (2010). The origin and history of soybean. In: Singh, G., editor, The Soybean: botany, production and uses. CABI, Wallingford, UK. p. 1–23

R Core Team. (2021). R: A language and environment for statistical computing.

Ravanel, S., Gakière, B., Job, D., & Douce, R. (1998). The specific features of methionine biosynthesis and metabolism in plants. *Proceedings of the National Academy of Sciences* 95(13): 7805 – 7812. https://doi.org/10.1073/pnas.95.13.7805.

Reinprecht, Y., Poysa, V.W., Yu, K., Rajcan, I., Ablett, G.R., & Pauls, K.P. (2006). Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (Glycine max (L.) Merrill) germplasm. *Genome* 49(12): 1510–1527. https://doi.org/10.1139/g06-112.

Roesler, K., Shen, B., Bermudez, E., Li, C., Hunt, J., Damude, H.G., Ripp, K.G., Everard, J.D., Booth, J.R., Castaneda, L., Feng, L., & Meyer, K. (2016). An improved variant of soybean type 1 diacylglycerol acyltransferase increases the oil content and decreases the soluble carbohydrate content of soybeans. *Plant Physiology* 171(2): 878–893. https://doi.org/10.1104/pp.16.00315.

Schuetzenmeister, A., & Dufey, F. (2020). VCA: Variance Component Analysis. *R package version 1.4.3.* R package. https://cran.r-project.org/package=VCA.

Serretti, C., Schapaugh, W.T., & Leffel, R.C. (1994). Amino Acid Profile of High Seed Protein Soybean. *Crop Science* 34: 207–209. https://doi.org/10.2135/cropsci1994.0011183X003400010037x.

Singer, W.M., Shea, Z., Yu, D., Huang, H., Mian, M.A.R., Shang, C., Rosso, M.L., Song, Q.J., & Zhang, B. (2022). Genome-Wide Association Study and Genomic Selection for Proteinogenic Methionine in Soybean Seeds. *Frontiers in Plant Science* 13. 2022 ttps://doi.org/10.3389/fpls.2022.859109.

USDA. (2023). World Supply and Use of Oilseeds and Oilseed Products. *Oil Crop Yearbook.* https://www.ers.usda.gov/data-products/oil-crops-yearbook/.

Vaughn, J.N., Nelson, R.L., Song, Q., Cregan, P.B., & Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3: Genes, Genomes, Genetics* 4(11): 2283–2294. https://doi.org/10.1534/g3.114.013433.

Vello, N.A., Fehr, W.R., & Bahrenfus, J.B. (1984). Genetic Variability and Agronomic Performance of Soybean Populations Developed from Plant Introductions 1 . *Crop Science* 24(3): 511–514. https://doi.org/10.2135/cropsci1984.0011183x002400030020x.

Warrington, C. V., Abdel-Haleem, H., Hyten, D.L., Cregan, P.B., Orf, J.H., Killam, A.S., Bajjalieh, N., Li, Z., & Boerma, H.R. (2015). QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. *Theoretical and Applied Genetics* 128(5): 839–850. https://doi.org/10.1007/s00122-015-2474-4.

Wickham, H. (2016). ggplot2. Elegant Graphics for Data Analysis. Springer-Verlag, New York, NY.

Wilcox, J.R., & Cavins, J.F. (1995). Backcrossing High Seed Protein to a Soybean Cultivar. *Crop Science* 35: 1036–1041. https://doi.org/10.2135/cropsci1995.0011183X003500040019x.

Yan, L., Hofmann, N., Li, S., Ferreira, M.E., Song, B., Jiang, G., Ren, S., Quigley, C., Fickus, E., Cregan, P., & Song, Q. (2017). Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics* 18(1): 1–11. https://doi.org/10.1186/s12864-017-3922-0.

Zarkadas, C.G., Gagnon, C., Poysa, V., Khanizadeh, S., Cober, E.R., Chang, V., & Gleddie, S. (2007). Protein quality and identification of the storage protein subunits of tofu and null soybean genotypes, using amino acid analysis, one- and two-dimensional gel

electrophoresis, and tandem mass spectrometry. *Food Research International* 40(1): 111–128. https://doi.org/10.1016/j.foodres.2006.08.005.

Zhang, Y., Li, W., Lin, Y., Zhang, L., Wang, C., & Xu, R. (2018). Construction of a high-density genetic map and mapping of QTLs for soybean (*Glycine max*) agronomic and seed quality traits by specific length amplified fragment sequencing. *BMC Genomics* 19(1): 1–14. https://doi.org/10.1186/s12864-018-5035-9.

Zhang, J., Song, Q., Cregan, P.B., & Jiang, G.L. (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theoretical and Applied Genetics* 129(1): 117–130. https://doi.org/10.1007/s00122-015-2614-x.

Zhao, X., Dong, H., Chang, H., Zhao, J., Teng, W., Qiu, L., Li, W., & Han, Y. (2019). Genome wide association mapping and candidate gene analysis for hundred seed weight in soybean [Glycine max (L.) Merrill]. *BMC Genomics* 20(1): 1–11. https://doi.org/10.1186/s12864-019-6009-2.

Table 3.1. Protein and sulfur-containing amino acid content of the RIL population parents.

Genotype	Protein	Cysteine	Cysteine/cp†	Methionine	Methionine/cp†	(Cys+Met)/cp†
	%					
Woodruff	38.79	0.68	1.75	0.55	1.42	3.17
PI 399000	44.16	0.81	1.83	0.63	1.43	3.26

Note: Analysis performed at the Experiment Station Chemical Laboratories, University of Missouri 2019.

† Normalized values of the amino acids are given as a percentage of the protein fraction (AA/Protein ×100).

Table 3.2. Variance components and heritability of seed composition and seed size.

	Protein	Oil	Cysteine (Cys)	Methionine (Met)	Cys+Met	Seed Size
Genotype	1.31***	0.49***	0.0006***	0.0004***	0.002***	3.00***
Environment	0.54***	0.03***	0.0007***	0.0005***	0.001***	1.07***
Replication	0.11***	0.02***	0.0001***	0.0001***	0.0003***	0.06***
G x E	0.30***	0.12***	0.0003***	0.0001***	0.0007***	0.61***
Residual	0.56	0.20	0.001	0.0006	0.003	0.69
Heritability	0.93	0.93	0.78	0.84	0.82	0.95
CV (%)	1.91	2.41	2.81	1.82	2.05	5.35

Note: Significance calculated with an ANOVA-Type Estimation of Mixed Models

*** Indicates the significance at a probability level of 0.001.

Table 3.3. Seed composition and seed size of the Woodruff × PI 399000 RIL population evaluated in 2020 and 2021.

Parent/ Population	Protein	Oil	Cysteine†	Methionine†	Cys+Met†	Seed Size
	%					g/100-seed
Woodruff	41.88	20.00	1.33	1.35	2.69	15.93
PI 399000	44.17	18.99	1.41	1.37	2.78	14.83
RIL Min.	36.96	15.87	1.24	1.26	2.54	9.30
RIL Mean	42.50	19.68	1.39	1.37	2.76	15.45
RIL Max.	46.99	22.98	1.55	1.48	2.99	23.80

Note: Protein, oil, cysteine, and methionine concentration are given as percentage of seed dry weight.

† Normalized values of the amino acids are given as a percentage of the protein fraction (AA/Protein ×100).

Table 3.4. Phenotypic correlation among the traits evaluated over six environments in the Woodruff × PI 399000 RIL population.

	Protein	Oil	Methionine (Met)	Cysteine (Cys)	Cys+Met
Oil	-0.50**				
Methionine	-0.47**	0.27**			
Cysteine	-0.37**	0.18**	0.46**		
Cys+Met	-0.48**	0.26**	0.79**	0.90**	
Seed Size	0.35**	-0.16**	-0.04*	-0.10**	-0.09**

*, ** Indicates significance at the 0.05 and 0.01 probability level.

Table 3.5. QTLs identified in the Woodruff × PI 399000 RIL population evaluated over six environments.

QTL name	Chr	Lod	# Env	Marker Interval	Wm82.a2.v1 interval (Mb)	PVE	Effect	Favorable Allele	Donor Parent
qPro-6	6	10.13	6	Gm06_9081030CA - **Gm06_9668798TC**	9.08 - 9.67	20.34	0.51	T	PI 399000
qPro-15	15	8.64	5	**Gm15_22112626GA** - Gm15_42869969CT	22.17 - 43.63	16.61	0.47	A	PI 399000
qPro-17	17	7.83	5	Gm17_7461701AC - **Gm17_7640951AG**	7.19 - 7.37	15.69	-0.45	A	Woodruff
qOil-14	14	10.82	6	Gm14_46205549CT - **Gm14_46635262TC**	45.48 - 45.92	20.35	0.31	C	PI 399000
qOil-17	17	8.91	6	Gm17_8109237AC - **Gm17_8270421AG**	7.84 - 8.00	14.96	0.27	G	PI 399000
qOil-10	10	6.13	3	Gm10_43178809GT - **Gm10_45051560TC**	43.76 - 45.63	12.39	0.25	C	PI 399000
qOil-6	6	5.30	4	Gm06_9081030CA - **Gm06_9442481TC**	9.08 - 9.45	11.11	-0.23	C	Woodruff
qOil-4	4	4.26	3	Gm04_7795928AG - **Gm04_8375095TC**	7.86 - 8.45	6.26	-0.17	T	Woodruff
qOil-19	19	4.99	2	**Gm19_46647127CT** - Gm19_48433644AG	46.76 - 48.56	7.40	0.19	C	PI 399000
qMet-3	3	5.23	2	**Gm03_5165511AC** - Gm03_13245008CT	5.07 - 13.24	11.26	0.01	A	PI 399000
qMet-6	6	7.68	4	Gm06_9081030CA - **Gm06_9668798TC**	9.08 - 9.68	15.96	-0.01	C	Woodruff
qMet-10	10	7.35	3	**Gm10_3022221TC** - Gm10_3962673CT	3.03 - 3.98	15.19	0.01	T	PI 399000
qMet-15	15	5.87	C†	Gm15_22112626GA - **Gm15_42869969CT**	22.16 - 43.63	11.75	-0.01	T	Woodruff
qCys-6	6	10.28	4	Gm06_10124470TC - **Gm06_10383065GA**	10.14 - 10.40	21.58	-0.01	C	Woodruff
qCys-10	10	9.52	5	**Gm10_3022221TC** - Gm10_3962673CT	3.03 - 3.98	20.88	0.01	T	PI 399000
qCys-15	15	5.75	C†	**Gm15_22112626GA** - Gm15_46930166AG	22.17 - 47.71	8.16	-0.01	G	Woodruff
qCM-3	3	4.15	3	**Gm03_5502496TC** - Gm03_13245008CT	5.40 - 13.24	7.91	0.01	C	PI 399000
qCM-6	6	10.29	5	**Gm06_10124470TC** - Gm06_10383065GA	10.14 - 10.40	20.59	-0.02	C	Woodruff
qCM-10	10	7.21	3	**Gm10_3022221TC** - Gm10_3962673CT	3.03 - 3.98	19.31	0.02	T	PI 399000
qCM-15	15	5.31	C†	**Gm15_22112626GA** - Gm15_46930166AG	22.17 - 47.71	11.54	-0.01	G	Woodruff
qSS-17	17	10.89	6	Gm17_7461701AC - **Gm17_7513295GA**	7.19 - 7.24	21.85	-0.81	A	Woodruff
qSS-20	20	7.23	6	**Gm20_40358501CT** - Gm20_41127680AC	41.47 - 42.24	14.80	-0.68	T	Woodruff

C† = QTL identified only in the combined analysis of all six environments.
Note: Most significant marker indicated in bold.

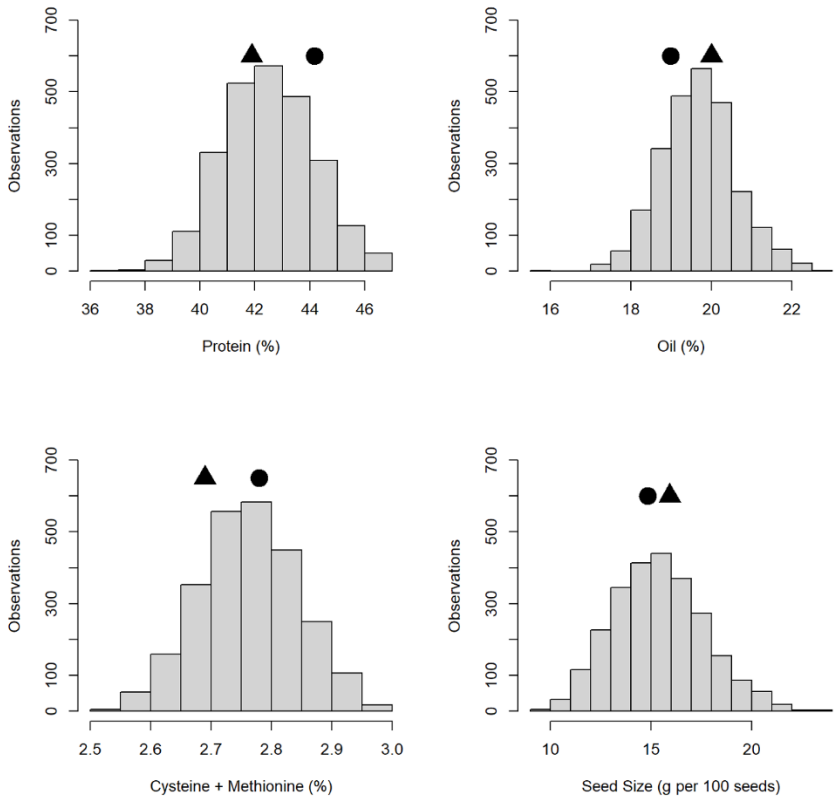

Figure 3.1. Distribution of protein, oil, cysteine + methionine and seed size evaluated across six environments. Black circle and triangle represent the values of PI 399000 and Woodruff, respectively.

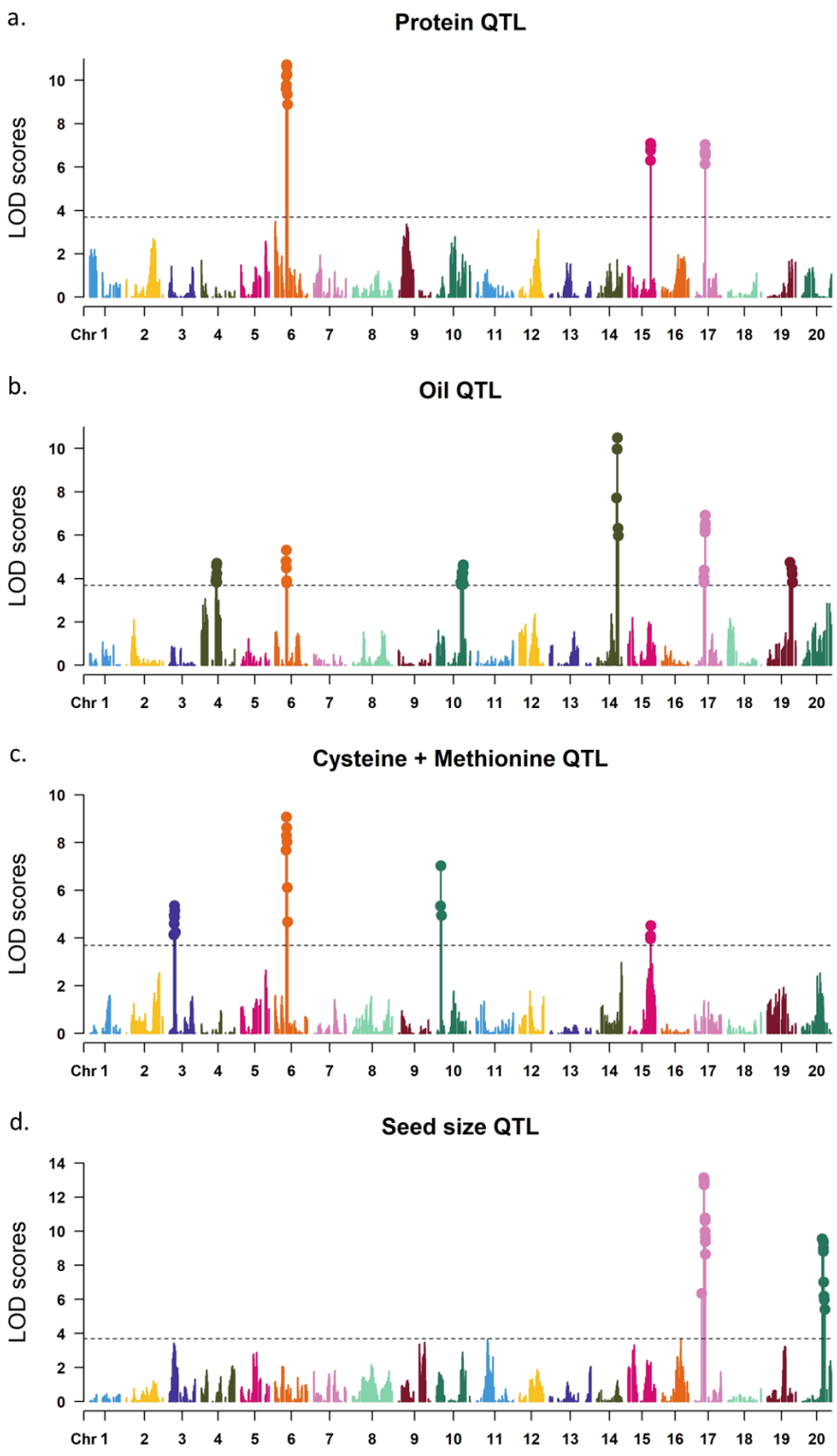

Figure 3.2. QTLs identified in the Woodruff × PI 399000 RIL population with the combined analysis of six environments. 3a) Protein; 3b) Oil; 3c) cysteine and methionine; and 3d) seed size.

Table 3.S1. Mean values of seed composition and seed size in the Woodruff × PI 399000 RIL population evaluated over six environments.

Environment	Protein	Oil	Cys + Met	Seed size	Temperature†	Precipitation‡
	%			g/100-seed	°C	mm
ATH20	42.52	19.82	2.82	16.09	23.2	721
ATH21	41.81	19.78	2.76	16.19	22.6	677
LSU20	43.74	19.52	2.74	16.59	24.1	473
LSU21	43.06	19.35	2.73	15.21	24.6	455
NC20	42.12	19.74	2.76	14.84	21.9	957
NC21	41.87	19.86	2.76	13.79	21.3	772

† Average temperature from June 01 to November 30.

‡ Total precipitation from June 01 to November 30.

Data retrieved from http://weather.uga.edu/, https://weather.lsuagcenter.com/, https://climate.ncsu.edu/office/ and https://www.weather.gov/

Table 3.S2. Summary of the linkage map constructed with 1865 markers in the Woodruff × PI 399000 RIL population.

Chromosome	Number of Markers	Length (cM)	Average spacing	Max spacing
1	76	122.95	1.64	24.09
2	115	150.03	1.32	19.39
3	82	106.37	1.31	9.97
4	74	136.44	1.87	19.44
5	90	113.77	1.28	18.95
6	122	131.79	1.09	12.22
7	80	130.48	1.65	15.88
8	140	161.98	1.17	8.08
9	91	130.80	1.45	22.98
10	94	137.64	1.48	16.10
11	81	149.36	1.87	11.05
12	81	98.83	1.24	14.23
13	109	167.87	1.55	31.05
14	86	102.10	1.20	8.44
15	88	111.56	1.28	12.64
16	85	109.42	1.30	9.43
17	74	108.61	1.49	10.06
18	123	136.57	1.12	17.12
19	98	116.37	1.20	13.33
20	76	120.58	1.61	9.89
Overall	1865	2543.52	1.38	31.05

Table 3.S3. Analysis of variance for the top two QTL effects for each trait evaluated.

	Df	Protein	Oil	Cys+Met	Seed size
QTL1	1	28.49***	17.69***	0.021***	127.07***
QTL2	1	24.03***	9.48***	0.046***	86.94***
QTL1 x QTL2	1	3.17*	0.26	0.00	1.93
Residuals	153	0.82	0.34	0.001	1.76

Note: For protein the two QTLs are qPro-6 and qPro-15. For oil, the QTLs are qOil-14 qOil-17. For Cys+Met, QTLs are qCM-6 and qCM-10. For seed size, QTLs are qSS-17 and qSS-20.

Table 3.S4. Candidate genes located within a 100 Kb window centered on the most significant marker for each QTL identified in five or more environments.

QTL	Chr	Gene	Start	Stop	PFAM protein family
qPro-6	6	Glyma.06G118300	9631662	9642200	RING/FYVE/PHD-type zinc finger family protein
	6	Glyma.06G118400	9639161	9639805	carbon/nitrogen insensitive 1
	6	Glyma.06G118500	9644661	9650144	Chalcone and stilbene synthase family protein
	6	Glyma.06G118800	9666921	9667708	spermidine hydroxycinnamoyl transferase
	6	Glyma.06G119100	9689732	9691834	gibberellin 20 oxidase 2
	6	Glyma.06G119200	9706114	9709403	wall-associated kinase
	6	Glyma.06G119400	9719172	9726230	S-adenosyl-L-methionine-dependent methyltransferase
qCM-6	6	Glyma.06G124200	10098710	10102918	Leucine-rich repeat transmembrane protein kinase protein
	6	Glyma.06G124300	10113576	10114445	Dof-type zinc finger DNA-binding family protein
	6	Glyma.06G124400	10128263	10129012	VQ motif-containing protein
	6	Glyma.06G124500	10136573	10142145	Galactosyltransferase family protein
	6	Glyma.06G124600	10142607	10146017	Ubiquitin-associated/translation elongation factor EF1B
	6	Glyma.06G124700	10156836	10160820	Leucine-rich repeat protein kinase family protein
	6	Glyma.06G124800	10164556	10167224	myosin heavy chain-related
	6	Glyma.06G124900	10168532	10171636	ubiquitin-conjugating enzyme 10
	6	Glyma.06G125100	10190469	10192862	cytokinin response factor 4
qCM-10	10	Glyma.10G034400	2981775	3003608	Myosin family protein with Dil domain
	10	Glyma.10G034600	3008811	3010158	C2H2 and C2HC zinc fingers superfamily protein
	10	Glyma.10G034800	3032154	3035186	BET1P/SFT1P-like protein 14A
	10	Glyma.10G034900	3036333	3041422	Outer membrane OMP85 family protein
	10	Glyma.10G035000	3045331	3046792	lateral organ boundaries-domain
	10	Glyma.10G035200	3069708	3073972	alpha/beta-Hydrolases superfamily protein
	10	Glyma.10g037100	3256351	3259176	glycinin G4
qOil-14	14	Glyma.14G193900	45884076	45890019	Clp-N motif-containing P-loop nucleoside triphosphate
	14	Glyma.14G194000	45892841	45893979	zinc ion binding
	14	Glyma.14G194100	45920719	45923885	zinc finger (CCCH-type) family protein
	14	Glyma.14G194300	45935668	45939896	fatty acid desaturase 8
	14	Glyma.14G194400	45951636	45954741	Pentatricopeptide repeat (PPR-like) superfamily protein
	14	Glyma.14G194500	45962394	45965333	SELT-like protein precursor
	14	Glyma.14G194600	45965417	45970808	alternative NAD(P)H dehydrogenase 1
	14	Glyma.14G194700	45971353	45977130	Ankyrin repeat family protein
qPro-15	15	Glyma.15G194900	22187694	22194246	endoplasmic reticulum oxidoreductins 2
	15	Glyma.15G049200	3874867	3876757	SWEET15, AtSWEET15|senescence-associated gene 29
qSS-17	17	Glyma.17G092400	7196181	7200287	spermidine synthase 1
	17	Glyma.17G092600	7220742	7225202	GATA type zinc finger transcription factor family protein
	17	Glyma.17G092700	7229709	7234743	Calmodulin binding protein-like
	17	Glyma.17G092800	7253511	7255621	Gibberellin-regulated family protein
	17	Glyma.17G092900	7257689	7264511	Glycosyl hydrolase family 47 protein
	17	Glyma.17G093000	7265450	7268134	DNA glycosylase superfamily protein
	17	Glyma.17G093300	7277777	7282818	S-adenosyl-L-methionine-dependent methyltransferase
	17	Glyma.17G093400	7287962	7289636	Transducin/WD40 repeat-like superfamily protein
qPro-17	17	Glyma.17G093900	7325884	7328325	response regulator 5
	17	Glyma.17G094000	7340914	7342597	RING/U-box superfamily protein
	17	Glyma.17G094100	7350712	7354261	ARM repeat superfamily protein
	17	Glyma.17G094300	7374357	7383061	plastid transcriptionally active 2
	17	Glyma.17G094400	7393359	7395553	Homeodomain-like superfamily protein
	17	Glyma.17G094700	7403909	7408160	serine/threonine protein kinase 2
	17	Glyma.17G094800	7410146	7431850	ARM repeat superfamily protein
qOil-17	17	Glyma.17G101300	7951772	7956449	AGAMOUS-like 65
	17	Glyma.17G101400	7956697	7960511	expansin A15
	17	Glyma.17G101500	7978357	7980301	NAC domain containing protein 100
	17	Glyma.17G101600	7987483	7989181	fucosyltransferase 1
	17	Glyma.17G101700	7993290	7997740	D-ribulose-5-phosphate-3-epimerase
	17	Glyma.17G101800	7998166	8001948	RING/U-box superfamily protein
	17	Glyma.17G101900	8004316	8005752	Tetratricopeptide repeat (TPR)-like superfamily protein
	17	Glyma.17G102100	8013190	8016414	Polynucleotidyl transferase, ribonuclease H-like protein
	17	Glyma.17G102200	8018190	8023260	CCT motif -containing response regulator protein
	17	Glyma.17G102300	8024107	8027410	Pentatricopeptide repeat (PPR) superfamily protein
	17	Glyma.17G102400	8027482	8029709	Galactosyl transferase GMA12/MNN10 family protein
	17	Glyma.17G102600	8034007	8036592	Protein kinase superfamily protein
qSS-20	20	Glyma.20G177200	41446962	41451980	cytosolic invertase 2
	20	Glyma.20G177300	41460432	41461592	glycine-rich protein
	20	Glyma.20G177600	41506046	41507720	LOB domain-containing protein 1
	20	Glyma.20G177700	41518436	41522203	ATP synthase D chain, mitochondrial
	20	Glyma.20G177800	41524383	41531869	mitochondrial substrate carrier family protein

Table 3.S5. Performance of top 13 RILs with protein content higher than 43% and cysteine + methionine higher than 2.8 %. Allele variation at each QTL identified is presented.

Line	Protein	Oil (%)	Cys+Met	Seed Size (g/100 seed)	*qPro06*	*qPro15*	*qPro17*	*qCM03*	*qCM06*	*qCM10*	*qCM15*	Protein QTL	CM QTL
Woodruff	42.27	20.69	2.74	16.63	CC	GG	AA	TT	CC	TT	GG	1	2
PI 399000	44.36	19.73	2.83	15.61	TT	AA	GG	CC	TT	CC	AA	2	2
G19-12017	43.38	20.35	2.88	16.87	CC	AA	AA	CC	CC	CC	AA	2	3
G19-11933	43.42	19.62	2.86	20.67	CC	AA	AA	CC	CC	CC	AA	2	3
G19-11963	43.55	20.33	2.83	15.89	TT	GG	AA	TT	TT	CC	GG	2	2
G19-11944	43.98	19.14	2.82	19.21	CC	AA	AA	CC	CC	CC	AA	2	3
G19-11889	43.54	20.08	2.82	16.97	TT	AA	AA	CC	TT	CC	AA	3	2
G19-12053	43.15	20.62	2.82	15.80	TT	AA	GG	CC	TT	CC	AA	2	2
G19-12013	43.64	20.47	2.82	17.36	TT	GG	AA	CC	TT	TC	GG	2	2
G19-11875	43.50	20.11	2.82	16.45	CC	AA	AA	TC	CC	CC	AA	2	2
G19-11867	43.69	19.91	2.81	18.44	TT	GG	AA	CC	TT	TT	GG	2	2
G19-12021	43.33	19.64	2.81	17.37	TT	GG	AA	CC	TT	TT	GG	2	2
G19-11897	43.58	19.62	2.81	15.48	TT	GG	AA	CC	TT	CC	GG	2	3
G19-11858	43.04	20.35	2.81	15.69	TT	GG	AA	CC	TT	CC	GG	2	3
G19-11930	44.21	19.88	2.80	18.82	TT	AA	AA	CC	TT	CC	AA	3	2

qPro-6 is from PI 399000 (Marker: Gm06_9668798_T_C); qPro-15 is from PI 399000 (Marker: Gm15_22112626_G_A); qPro-17 is from Woodruff (Marker: Gm17_7640951_A_G); qCM-3 is from PI 399000 (Marker: Gm03_5502496_T_C); qCM-6 is from Woodruff (Marker Gm06_10124470_T_C); qCM-10 is from PI 399000 (Marker: Gm10_3066211_C_T) and qCM-15 is from Woodruff (Marker: Gm15_22112626_G_A).

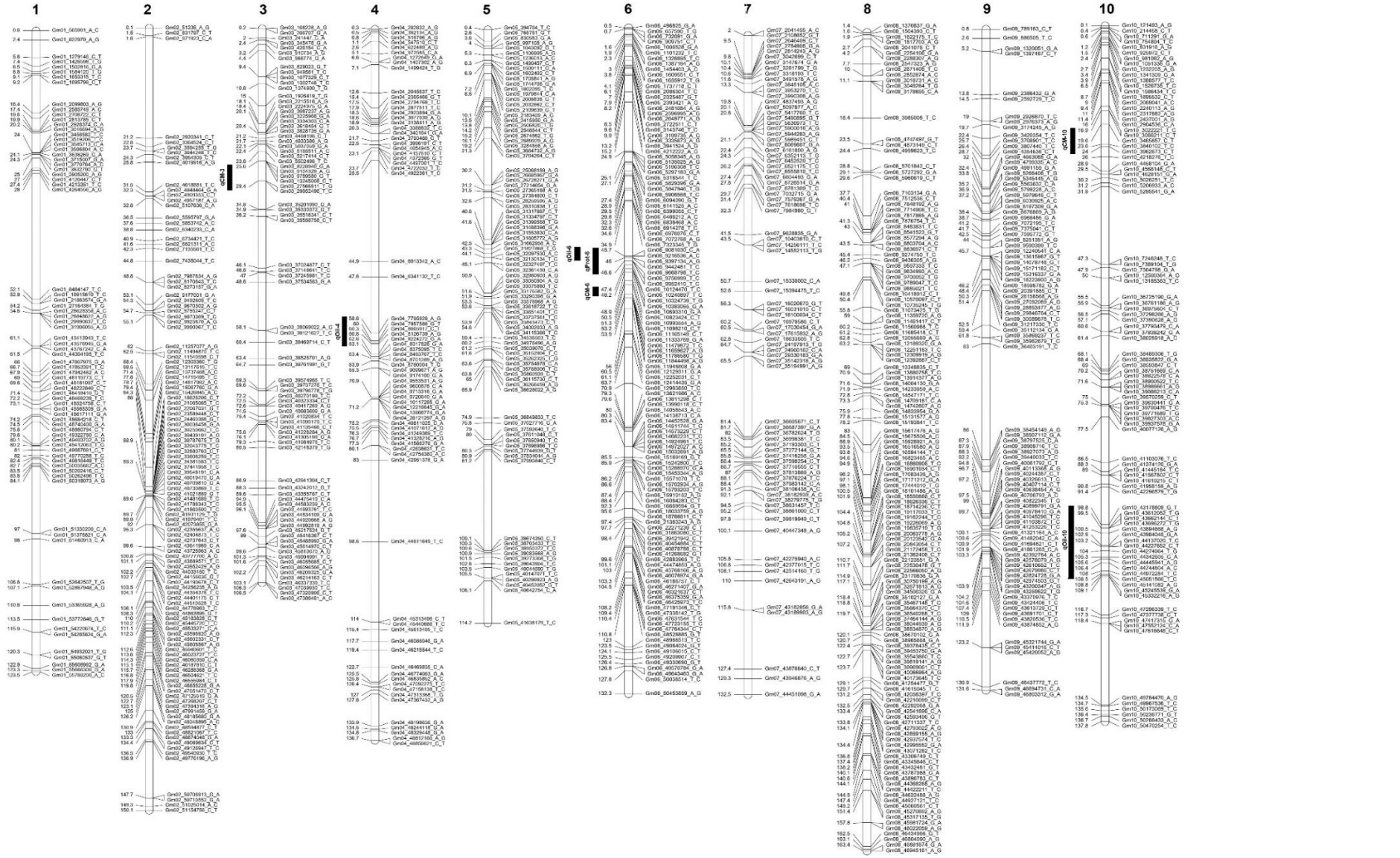

Figure 3.S1. Genetic map of the Woodruff × PI 399000 RIL population Chr 1 to Chr 10.

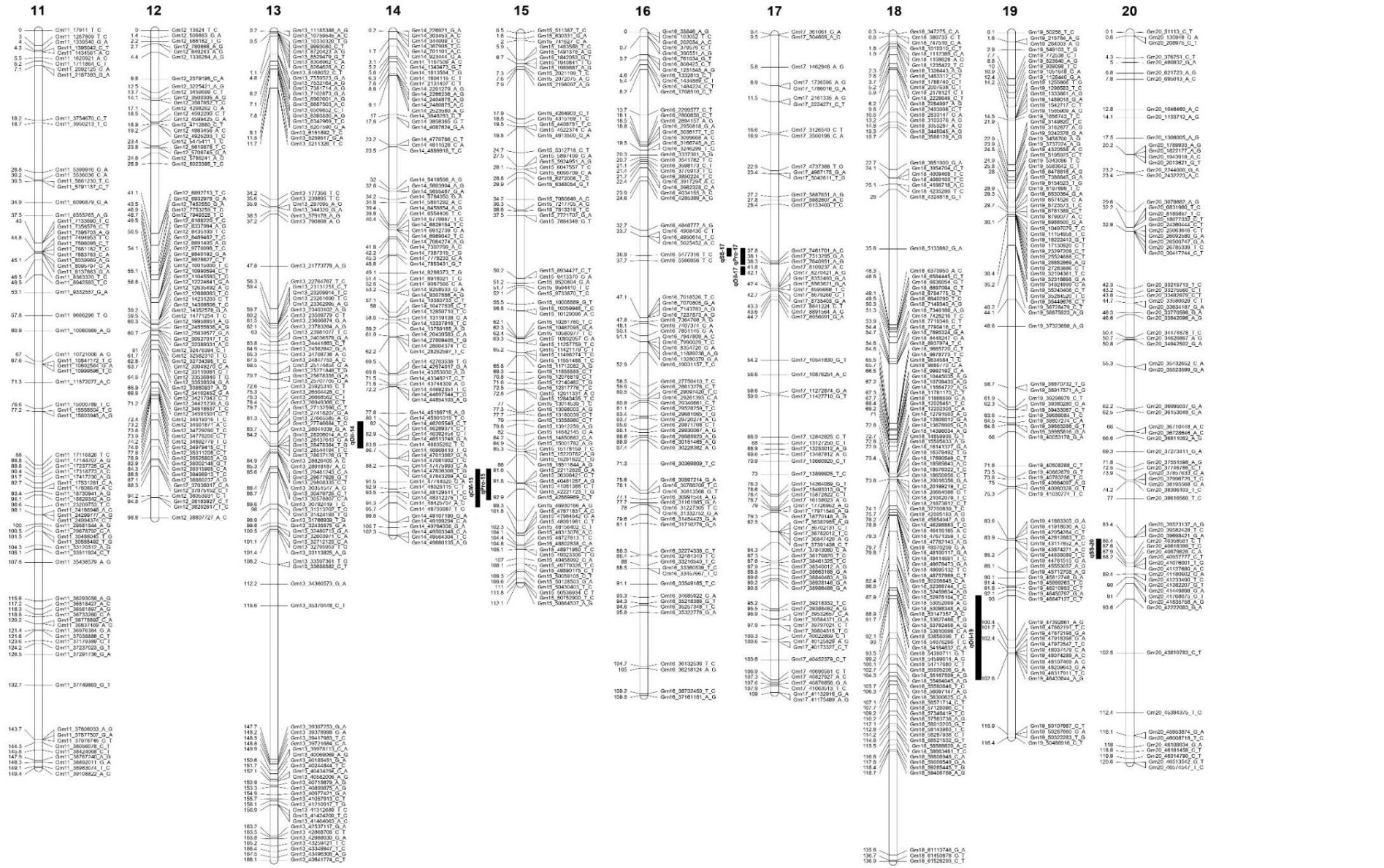

Figure 3.S2. Genetic map of the Woodruff × PI 399000 RIL population Chr 11 to Chr 20

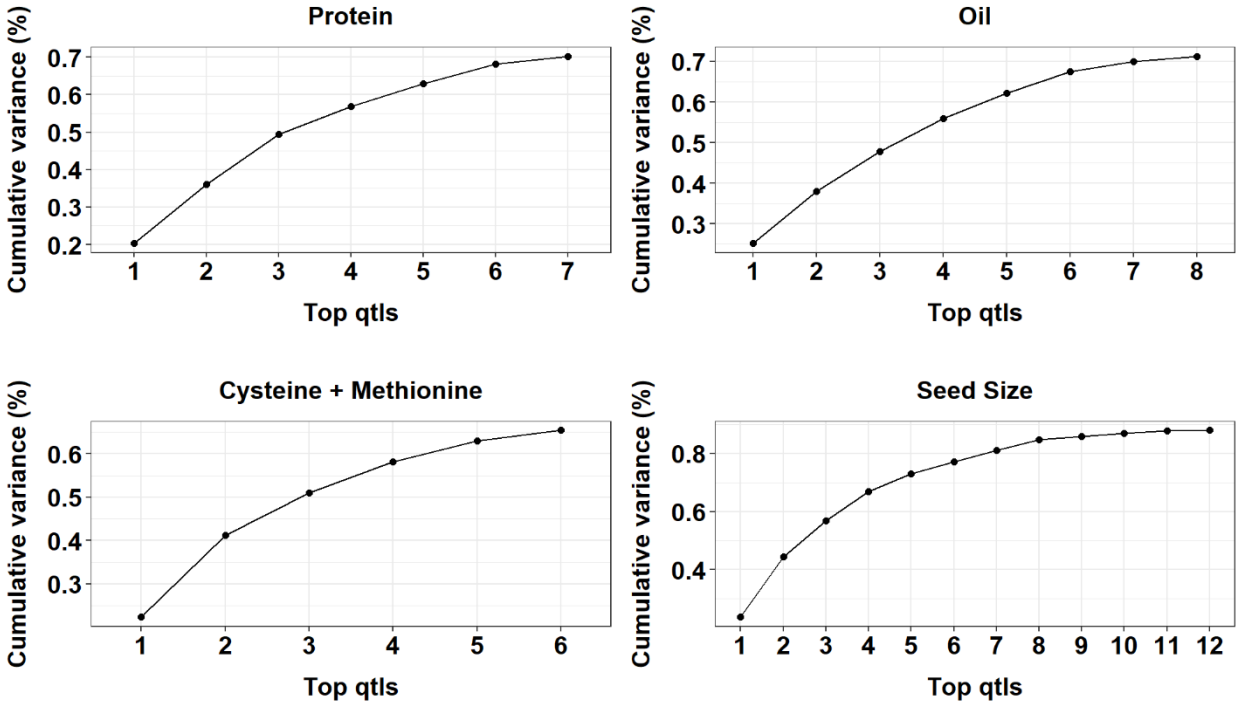

Figure 3.S3. Cumulative variance of QTLs with LOD>2.5 for protein, oil, cysteine + methionine and seed size, respectively.

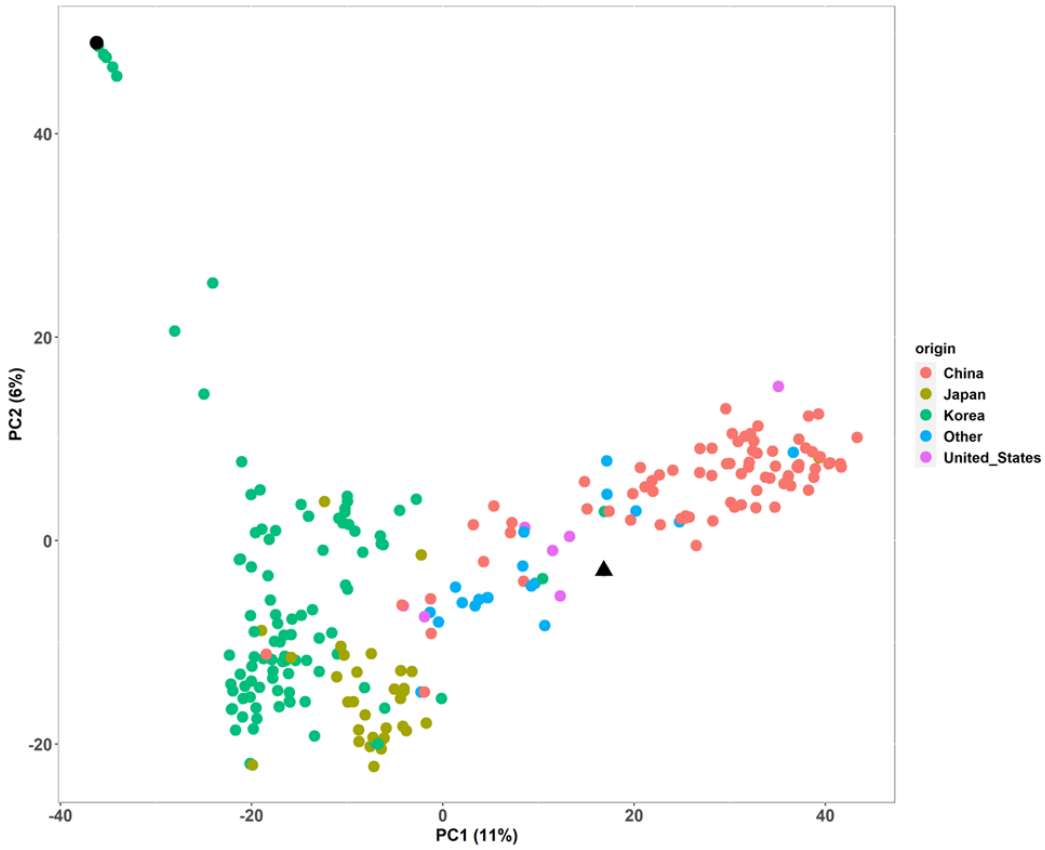

Figure 3.S4. Principal Component Analysis of 254 accessions from the USDA Soybean Germplasm Collection, with protein > 44%, Cys + Met > 3% and Maturity Group > V. Analysis performed with 24,424 SNPs from the SoySNP50K. Black circle and triangle represent the PI 399000 and Woodruff, respectively.

CHAPTER 4

SUMMARY

Soybean is one of the most important crops. It is the main source of protein for livestock and provides oil for cooking and industrial uses. The quality of the soybean meal is directly related to the seed protein content and other composition traits. To ensure that the quality parameters are met, breeding for improved seed composition is needed. In this context, it is important to understand the genetic control of the protein and amino acids and their relationship with other traits such as yield and oil content.

In chapter 2, the genetic control of soybean protein composition, the stability across different populations and the relationship of protein with other traits was studied. Previously, a QTL on Chr 20 from Danbaekkong has been demonstrated to be associated with elevated protein content. In the present study, introgression of the Chr 20 QTL from Danbaekkong was performed and the QTL was introgressed into a wide range of genetic backgrounds genetic improvement of protein content. The high protein allele was evaluated in a multiparent population through multiple years. Based on the previously publications, the TaqMan marker GSM1252 was developed to track the high protein allele and understand its effects. The high protein allele increased the seed protein content by 3.3% on average with independence of the population backgrounds, with values ranging from 2.7 to 3.7%.The impact of the high protein allele on yield was also evaluated in six environments, indicating a trend of reduction in yield as protein level increases. However, for the majority of the populations tested was no significant difference ($p<0.05$) in yields between lines with high protein and normal protein content. Overall, the results demonstrated that the high

protein Danbaekkong allele was highly stable across different genetic backgrounds and the negative effects of the high protein phenotypes on yield can be mitigated through selection. The marker GSM1252 was used to genotype 35 *Glycine max* ancestors of North American soybean cultivars and 79 *Glycine soja* core accessions from the USDA Soybean Germplasm Collection. The results indicated that all 35 *G. max* accessions carry the low protein allele variant, while the *G. soja* core accessions possess the high protein allele. The result also revealed that the Danbaekkong Chr 20 high protein allele originated from a *G. soja* accession (PI 163453) and there is an opportunity to use QTLs from *G. soja* to increase the levels of protein in soybean breeding.

In chapter 3, the genetic architecture of the amino acid composition of the protein was studied using a bi-parental population. A RIL population derived from a cross of Woodruff × PI 399000 was evaluated in six environments and genotyped with the SoySNP6K. QTL analysis was performed with 1865 SNPs and three loci were identified on Chrs 6, 15 and 17 across all environments for protein, six QTLs on Chrs 4, 6, 10, 14, 17, and 19 in at least two environments for oil, three QTLs Chrs 3, 6 and 10 for cysteine and methionine and two QTLs on Chrs 17 and 20 for seed size in all environments. Combination of favorable QTL alleles inherited from PI 399000 can increase protein content without decreasing concentration of cysteine + methionine, which usually have a negative correlation with high protein. The markers linked to the QTLs for high protein and elevated amino acid concentration can be used to breed soybean cultivars with improved soybean meal.