Fine Mapping and Candidate Gene Identification of a Soybean Seed Protein and Oil QTL from a Wild Soybean Accession and Linkage Analysis for Whole Plant Biomass, Carbon, Nitrogen, and Seed Composition using a RIL Mapping Population

---

A Thesis

presented to the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment of the Requirements for the Degree of Master of Science in

Plant Breeding, Genetics, and Genomics

---

by

Yia Yang

Dr. Andrew Scaboo and Dr. Jason Gillman, Thesis Supervisors

The undersigned, appointed by the Dean of the Graduate School,

have examined the thesis entitled

Fine Mapping and Candidate Gene Identification of a

Soybean Seed Protein and Oil QTL from a Wild Soybean Accession and

Linkage Analysis for Whole Plant Biomass, Carbon, Nitrogen, and Seed Composition using a RIL Mapping Population

Presented by YIA YANG

A candidate for the degree of Master of Science in Plant Breeding, Genetics, and Genomics,

and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Dr. Andrew M. Scaboo, Chair

_____

Dr. Jason D. Gillman, Co-Chair

_____

Dr. Trupti Joshi, Committee member

ACKNOWLEGEMENTS

I would like to thank my advisors Dr. Andrew M. Scaboo and Dr. Jason D. Gillman for their supportive guidance and for this great opportunity. I know that it was not easy for everyone around the world during the global COVID-19 pandemic. Their patience, kind words, wisdom, helped enabled me to pursue a Master of Science degree at the University of Missouri.

I would like to thank Dr. Trupti Joshi for accepting to be on my thesis committee. Her wisdom and supportive nature helped guide me through my writing by correcting my mistakes. I would like to thank her research lab and graduate student for helping me analyze data.

Thank you to the current and past research members of the University of Missouri Northern Soybean Breeding team for assisting me in my field experiments and lab work. They are a hard-working group that I consider not just as co-workers, but as lifelong friends.

Also, I would like to thank Dr. Jason D. Gillman's USDA-ARS genetics research lab members for assisting me in field experiments, greenhouse experiments, and lab work. Their work helped lessen the workload on my shoulders.

I would like to express my gratitude to the USDA-ARS, Missouri Soybean Merchandise Council and the United Soybean Board for their financial support for my research and graduate studies.

Finally, I would like to thank my family and friends for their support during my graduate studies. Their love, supportive words, and sacrifices enabled me to pursue my goals.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

Soybean [*Glycine max* (L.) Merr] cultivars have low genetic variation due to domestication, founder events, and selection strategies for modern plant breeding. There is a need to introduce genetic diversity into soybean cultivars for long-term improvement of agronomic and seed compositional traits. In both public and private soybean breeding programs, the introgression of wild soybean (*Glycine soja* Siebold & Zucc.) genes has been utilized to incorporate novel genetic diversity. In our study, 3,015 single $F_{4:9}$ soybean plants were genotyped for nine genotype-by-sequencing markers from a previous genetic mapping study on recombinant inbred lines (La, 2018) to create two residual heterozygotes derived near isogenic lines (RHD-NIL) populations. The first RHD-NIL population was selected for a novel oil quantitative trait loci (QTL) on chromosome 8 and the second RHD-NIL population was selected for a novel protein QTL on chromosome 14. Both novel QTL derived from the wild soybean accession PI 593983. The objective of this research is to validate these QTL, reduce the QTL interval, and fine map the two novel QTL for candidate gene identification. Single marker analysis and linkage analysis was conducted using SoySNP6K BeadChip markers for QTL validation. The chromosome 8 oil QTL was not advanced for fine mapping because the QTL was not validated in a subsequent field and greenhouse study. Whole genome resequencing was leveraged to reduce the QTL from 16.5 Mbp to approximately 4.6 Mbp and to fine map 50 high protein RHD-NIL, which have segregated for the validated chromosome 14 QTL to permit candidate gene identification. A total of 55 potential candidates was identified in a physical interval of 8,059,955 to 12,648,760 bp. Our results provide a better insight of utilizing wild soybean as a source of genetic diversity for soybean cultivar improvement.

In addition to the fine mapping and candidate gene identification study, we conducted linkage analysis for a recombinant inbred line (RIL) mapping population for plant biomass content, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content. Soybean seeds require a large amount of nitrogen because of its high protein content. Through a symbiotic association between soil microorganisms and soybean root nodules, soybean is able to fix atmospheric dinitrogen for nitrogen uptake. Plant biomass was collected by bulking five soybean shoot samples per plot from 262 plots in four locations and bulking three soybean shoots samples per plot from 262 plots in one location. Plant materials were dried and weighed for whole plant biomass weight. Whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content was analyzed via near infrared spectroscopy. The objective of this study was to examine nitrogen mobilization from a mapping population from the cross PI 361103 (contains high shoot N content and low seed N content) x PI 567572B (contains high seed N content and low shoot N content), identify QTL for plant biomass, whole plant carbon content, whole plant nitrogen content, and seed composition, and study maternal effects of cytoplasmic inheritance of the five traits from the reciprocal parental cross. Linkage analysis was conducted using BARCSoySNP50K markers. We identified six QTL for plant biomass, two QTL for whole plant carbon content, three QTL for whole plant nitrogen content, three QTL for seed oil content, and five QTL for seed protein content, with multiple traits having overlapping QTL intervals. Our results indicate QTL associated with multiple traits demonstrating the potential of pleiotropic effect in our mapping population.

Chapter I

Introduction and Literature Review

Soybean Production and Products

Soybean [*Glycine max* (L.) Merr] is one of the most valuable crops in the world due to the high content and quality of seed protein and seed oil, which have uses as feed for livestock, a good source of protein and oil for human health, and as a biofuel stock (Masuda et al., 2009). The world total soybean production in 2019 was approximately 334 million metric tons (FAOSTAT, 2020). Soybean seeds contain approximately 20% oil and 40% protein on a dry weight basis, and the two most important economical components of soybean are its oil and soybean meal (Wilson, 2004; Warrington et al., 2015).

The three leading countries for soybean production in 2019 were Brazil, the United States, and Argentina, which produced 80% of the world's soybeans at 266 million tons metric (FAOSTAT, 2020). Brazil overtook the United States as the leading country in soybean production in 2019 by producing 114 million metric tons (FAOSTAT, 2020). The United States produced 97 million metric tons, which is 23 million metric tons less than the previous year and Argentina produced 55 million metric tons (FAOSTAT, 2020). The decrease in soybean production in the Unites States was due to 7.8 million hectares left unplanted because of flood and heavy precipitation damages (Clayton, 2019).

The increased demand of soybean meal as a protein source in animal feed has resulted in an increase in soybean production (Dei, 2011). Soybean meal is important for animals such as poultry, swine, and beef. Fifty-three percent of soybean meal sold in the United States was used in feed for poultry, 29% for swine feed, 8% for aquaculture, 7% for other animals, 2% for dairy,

<1% for cattle feed, and <1% for companion animals (USB, 2019) (Figure 1-1). Soybean oil is used as cooking oil, mayonnaise, salad dressing, and can also be used for industrial materials such as cement components, construction materials, electrical insulation, plastic, paint, mineral oils, and numerous other applications (Hammond et al, 2005).

Genetic Diversity

Current U.S. soybean [*Glycine max (L.)* Merr.] cultivars have relatively low genomic variation which has the potential to restrict genetic improvement for grain yield, seed quality, and other agronomic and quality traits (Hyten et al., 2006). Soybean is an autogamous species which leads to homozygosity and decreased genetic diversity (Hyten et al., 2006). Evolutionary events such as domestication, founder events, and selection can create genetic bottlenecks that can decrease genetic diversity, shift allelic frequencies, increase linkage disequilibrium (LD), and eliminate rare alleles (Halliburton, 2004). Founder events occurs when a few individuals are used to introduce a crop into a new region, or a few cultivars are used for breeding and selection aimed at crop improvement (Hyten et al., 2006). Domestication is when human's create cultivar species by artificially selecting wild species for crop improvement (Harlan et al., 1973). Hyten et al. (2006) reported that 50% of the genetic diversity and 81% of the rare alleles have been lost during domestication and that 60% of the genes show significant changes in allelic frequency.

In public soybean breeding programs, the introgression of wild soybean (*Glycine soja* Siebold & Zucc.) genes has been utilized to incorporate novel genetic diversity (Akpertey et al., 2014; Pratap et al., 2021). Modern soybean cultivars [*Glycine max (L.)* Merr.] originated from Southeast Asia and were first domesticated from wild soybean (*Glycine soja* Siebold & Zucc.) in

China, Japan, and Korea before spreading to the Americas (Carter et al., 2004; Wilson, 2008). Wild soybean is considered to be the progenitor of cultivated soybean; they both have 20 chromosomes (2n = 40) and can be sexually crossed which carries out normal meiotic chromosome pairing, which will produce fertile hybrids (Carter et al., 2004). For these reasons, wild soybean is a valuable resource for introgression of novel genes and for soybean cultivar development (Stupar, 2010). *G. soja* and *G. max* are morphologically distinct in that *G. soja* generally flowers later, lodges more, has more lateral branches, produces small black seeds, and tends to shatter its seeds (Liu et al., 2007).

Hyten et al. (2006) studied four groupings, *G. soja* , *G. max* Asian landraces, North American ancestor lines, and *G. max* elite cultivars on the impact of genetic bottlenecks. Among the four studied groupings, *G. soja* had the greatest number of observed unique variation with 237 single nucleotide polymorphisms (SNP) identified (Figure 1-2). Nucleotide diversity or sequence variation are measured in the unit π (pi) (Hyten et al., 2006). Hyten et al, (2006) reported the decrease of nucleotide diversity (π) from 2.17×10^{-3} in wild soybeans to 1.47×10^{-3} in landraces, to 1.14×10^{-3} in North American ancestors, and to 1.11×10^{-3} in elite cultivars, which indicated the bottleneck effects in soybean domestication. Zhou et al. (2015) resequenced 302 wild and cultivated accessions, which included 93 diverse accessions previously examined by Hyten et al. (2006) and reported that genetic diversity (π) has decreased from 2.94×10^{-3} in wild soybean to 1.40×10^{-3} in landraces and to 1.05×10^{-3} in soybean cultivars. Li et al. (2013) and Valliyodan et al. (2016) reported similar declines in nucleotide diversity. These results were consistent with Wang and Takahata (2007), which reported the genetic diversity index (π) of the Chinese and Japanese population at 0.809 and 0.814, respectively.

La et al. (2019) characterized agronomic and seed composition traits in a genetically diverse core collection of *G. soja* plant introductions (PI). Song et al. (2015) analyzed all of the *G. soja* PIs from the USDA soybean collection to show a total of 806 *G. soja* accessions from China, Korea, Japan, and Russia were nonredundant. La et al. (2019) selected ~10% of the 806 nonredundant *G. soja* accessions for a total of 80 *G. soja* PI which comprise a mini-core collection based on SNP diversity and genetic distance. The mini-core collection was a representation of the entire USDA collection of *G. soja* entries (Figure 1-3). Total protein content for the core collection ranged from 396.2 – 481.7 kg^{-1}, whereas the total oil content ranged from 157.6 – 175.8 kg^{-1} on a dry weight basis.

The strong pressure of domestication, modern plant breeding, and founder events have caused a decrease in overall genomic variation in cultivated soybean (Halliburton, 2004). Selection for increase grain yield has caused seed compositional traits such as protein content to decrease (Rincker et al., 2014; Boehm et al., 2019). Rincker et al. (2014) evaluated genetic improvement of US soybean in maturity groups II, III, and IV and reported an estimated linear rate of genetic yield gain of 23 kg ha^{-1} yr^{-1} in both maturity group II and maturity group III, and 20 kg ha^{-1} yr^{-1} in maturity group IV which resulted from field tests that revealed seed yield consistently increased over the past 80 years (1925 – 2005) due to breeding efforts (Figure 1-4). Boehm et al. (2019) had a similar study but focused on maturity groups V, VI, and VII from 1930 - 2010 and reported a linear rate of genetic yield improvement of 17.6 kg ha^{-1} yr^{-1} for maturity group V, 13.5 kg ha^{-1} yr^{-1} for maturity group VI, and 10.3 kg ha^{-1} yr^{-1} for maturity group VII (Table 1-1). Seed protein concentration decreased at a rate of 0.22 g kg^{-1} yr^{-1} in both maturity group II and maturity group III, and 0.16 g kg^{-1} yr^{-1} in maturity group IV (Rincker et al., 2014) (Figure 1-5). Maturity group VI soybeans saw a similar trend as maturity group II and III

with a protein content decrease rate of 0.23 g kg^{-1} yr^{-1} and maturity group VII had a decrease rate of 0.09 g kg^{-1} yr^{-1}, while maturity group V had an increase rate of 0.02 g kg^{-1} yr^{-1} (Boehm et al., 2019) (Table 1-1). Inversely, seed oil concentration increased at a rate 0.14 g kg^{-1} yr^{-1}, 0.10 g kg^{-1} yr^{-1}, 0.05 g kg^{-1} yr^{-1}, 0.07 g kg^{-1} yr^{-1}, 0.04 g kg^{-1} yr^{-1}, and decreased at a rate of 0.03 g kg^{-1} yr^{-1} in the maturity groups I, II, III, IV, VI, VII, and V, respectively (Rincker et al., 2014; Boehm et al., 2019). Simultaneous breeding for higher seed protein, seed oil, and yield in soybean germplasm can be difficult due to the negative correlation between seed protein and yield, seed protein and seed oil, and the positive correlation between seed oil and yield (Rincker et al., 2014, Wilson, 2004). These general correlation trends have been reported numerous times in the literature. Seed yield in both studies was reported to have increased considerably mainly due high selection. Specht et al. (2014) also concluded that the increase of seed yield for on-farm genetic gains in the US was due to genetic improvements, such as modern plant selection and breeding methods and introgression of resistance and tolerant pest traits, and agronomic improvements, such as precision agriculture and improved agricultural equipment.

Agronomic traits were also altered in response to breeding, for example plant lodging scores decreased over time, ranging from 0.01 score yr^{-1} – 0.02 score yr^{-1} in maturity groups II, II, and III (Rincker et al., 2014) and a decreased range of 0.009 score yr^{-1} – 0.01 score yr^{-1} in maturity groups V, VI, VII (Boehm et al., 2019). In all maturity groups, lodging improved ~1.0 on lodging scale score, which indicates that lodging is a trait selected by plant breeders (Rincker et al., 2014; Boehm et al., 2019). Seed size in both studies saw similar results with maturity groups II, VI, and VII having a decreased rate of 0.01 g yr^{-1} – 0.02 g yr^{-1}. While maturity groups III, IV, V saw an increase rate of seed size from 0.02 g yr^{-1} – 0.01 g yr^{-1} (Rincker et al., 2014; Boehm et al., 2019). Seed size was not significantly different in both studies, and researchers

concluded that seed size was not a trait selected by plant breeders (Rincker et al., 2014; Boehm et al., 2019). Days to maturity was measured with September 1st being the first maturity date (Rincker et al., 2014; Boehm et al., 2019). Rincker et al. (2015) reported a linear increase rate of 0.9 – 0.1 d yr^{-1} in maturity groups I, II, and III. While Boehm et al. (2019) reported an increase rate of 0.05 d yr^{-1} in maturity group V and a decrease rate of 0.002 – 0.02 d yr^{-1} in maturity groups VI and VII. Newer maturity group V cultivars are reaching maturity faster which is resulting in southern US plant breeders integrating maturity groups III and IV into their breeding program (Boehm et al., 2019).

The United States Department of Agriculture (USDA), Soybean Germplasm Collection maintains a collection of 21,810 accessions of the genus Glycine with 19,626 being cultivated soybean and 1,179 wild soybeans (USDA Soybean Germplasm Collection, accessed on May 20, 2020). Gizlice et al. (1994) studied 258 publicly developed cultivars pedigrees between 1947 – 1988 and reported that greater than 84% of their parentage can be traced to 17 ancestral lines. Publicly released cultivars in the US are derived from approximately 80 ancestral lines and most originated from China (Gizlice et al., 1994). Approximately there are 8,500 *G. Soja* accessions and 45,000 *G. max* Asian landraces accessions in storage around the world (Hyten et al., 2006; Wen et al., 2009).

When crossing *G. soja* with *G. max*, undesirable traits are often present in direct progeny from *G. soja* such as late flowering, hard seed coat, poor lodging, small seed size, pod shattering, and black color seeds (Carter et al., 2004, Liu et al., 2007). Many desirable genes from *G. soja* are thought to be linked to undesirable traits, making breeding with *G. soja* both time consuming and resource intensive (Carter et al., 2004). Marker-assisted selection may have the utility in breaking desirable traits with undesirable traits during successive backcrossing (Concibdi et al.,

2003). Rare alleles are often lost during domestication or due to founder events. Such alleles have largely untapped potential for soybean improvement (Hyten et al., 2006). There is the potential to address the underlying issues with low genetic diversity in North American public soybean breeding programs through introgression of alleles from wild soybean. Many previous studies have demonstrated the potential of alleles from wild soybean or crosses between *G. soja* and *G. max* for genetic and agronomic improvements, as well as being a source to identify new genes and alleles (Nawaz et al., 2018). Sundaramoorthy et al. (2016) and Kim et al. (2017) reported multiple loci from wild soybeans that controls flower color. Soybean cyst nematode (SCN) is one of the biggest threats to soybean productivity globally. Zhang et al. (2017) discovered SNP and candidate genes that are significantly associated with SCN with the use of wild soybeans. Many QTL related to wild soybean including yield and maturity (Li et al., 2008), SCN (Zhang et al., 2017), seed yield (Concibido et al., 2003), linolenic acid content (Pantalone et al., 1997), seed protein content (Diers et al., 1992), and salt tolerance (Ha et al., 2013 and Qi et al., 2014) have been mapped and identified from wild soybeans or from populations developed from *G. soja* and *G. max* crosses. Beche et al. (2020) conducted a linkage analysis and used a nested association mapping panel consisting of 392 F_4-derived recombinant inbred lines (RIL), developed from three biparental cross-combination of *G. max* and *G. soja*. Beche et al. (2020) reported a novel grain yield QTL (qGY-17) which showed a 6% increase in grain yield for the *G. soja* allele when compared to the *G. max* allele across all RIL and environments. Thus, wild soybean can be a powerful source of genetic diversity to improve soybean cultivars.

Mapping Populations for Identifying QTL

The most common types of mapping populations in plants used for quantitative traits analysis are RIL and near-isogenic lines (NIL). RIL are formed by crossing genetically divergent parental lines followed by repeated selfing or sibling mating (Broman, 2004; Pollard, 2012). RIL can be mapped to a casual QTL from their phenotypes varying genetically from one another (Pollard, 2012). The advantages of RIL are that their level of heterozygosity per locus are halved every generation, thus increasing their level of homozygosity and variation in genotypes (Blanco et al., 1998). A sufficiently advanced RIL population only needs to be genotyped for segregating markers once and through inbreeding, recombination frequency is increased between two closely linked markers beyond that possible for an F_2 population (Blanco et al., 1998). Environmental influences on quantitative traits can be reduced by evaluating multiple plants with the same genotype and at multiple locations (Blanco et al., 1998). The recombination frequency in a RIL population is higher than in an equally sized NIL population because of the accuracy of QTL localization, mapping resolution, population size, and RIL require less individuals, making RIL great for defining QTL regions (Keurentjes et al., 2007). Linkage analysis or QTL mapping requires a creation of a genome-wide linkage map and relies on markers being close to the casual loci to show a nonrandom association with the phenotype (Blanco et al., 1998; Pollard, 2012).

A genome-wide association study (GWAS) is based on linkage disequilibrium and is a powerful genetic tool to detect genetic variations of quantitative traits by using genome-wide markers combined with phenotypes (Zhang et al., 2019). GWAS does not require a genome-wide linkage map of a mapping population and can simultaneously analyze multiple alleles from the same locus, which provides a higher resolution but due to genetic relationships, it may also lead to multiple false positive results (Zhang et al., 2019).

NIL are important in molecular breeding to efficiently identify genes associated with the trait of interest as they contain identical genetic makeups except for few specific locations or genetic loci (Muehlbauer et al., 1988; Young et al., 1988; Keurentjes et al., 2006). True NIL are thought to contain only a single introgression per line, which increases the power to detect small-effect QTL (Keurentjes et al., 2007). NIL have been used in many studies in identifying targeted genes in many crops and species (Yuan et al., 2017). The use of NIL does not allow testing for genetic interactions, epistasis, and thereby minimizes the effects caused by different genetic backgrounds (Keurentjes et al., 2007; Yuan et al., 2017). NIL can be obtained by backcrossing a genotype containing an allele of interest to a genotype containing the desired background alleles (Keurentjes et al., 2007). The progenies are selected to retain alleles of interest and go through repeated backcrossing and extensive genotyping (Muehlbauer et al., 1988) or advancing generations of RIL by selfing (Brechenmacher et al., 2015; Glover et al., 2004; Kim et al., 2011; Yuan et al., 2017). NIL are useful for QTL validation and confirmation because they are developed to segregate for QTL in an otherwise homogeneous background (Glover et al., 2004). NIL are the preferred mapping population when wild and cultivated germplasms are combined (Eshed and Zamir, 1995; Jeuken and Lindhout; 2004; Von Korff et al., 2004).

Previous studies have demonstrated the use of NIL for QTL confirmation (Glover et al., 2004). Glover et al. (2004) studied a population from a cross between the cultivar Bell (Nickell et al., 190) and Colfax (Graef et al., 1994). The NIL were developed from F_4 derived lines and then advanced to F_7 as bulks and selected individual F_7 plants were threshed. The F_4 derived lines were heterozygous on chromosome (Chr.) 16 which carries the SCN resistance QTL. There are three NIL populations with NIL population 1 and 2 (NIL1, NIL2) having 48 lines and population 3 (NIL3) having 56 lines. NIL1 and NIL3 are predicted to be homozygous for the

susceptible allele at *rhg1* and NIL2 is predicted to be homozygous for the resistance allele at *rhg1*. The original F_4 population and the NIL populations were evaluated for resistance to SCN populations PA3 (HG type 7, race 3) and PA14 (HG type 1.3.5.6.7, race 14). Glover et al. (2004) identified regions on Chr. 6, 16, and 18 which were significantly associated with resistance to PA3 and PA14 by single-factor analysis. Seven markers mapped near the QTL on Chr. 16, two markers mapped near *rhg1*, and the 6 markers mapped to six other chromosomes. From a three-factor analysis, Glover et al. (2004) concluded that markers Satt431 and Satt277 together explained 87% of the phenotypic variance for PA3 and 64% of the phenotypic variance for PA14. Satt431 was significantly associated with resistance in each of the two tests with PA14 across all three NIL populations. Satt431 showed greater resistance for the Bell allele than lines homozygous for the Colfax allele in all the populations. Glover et al. (2004) confirmed the QTL on Chr. 16 for PA14 resistance from Bell and the QTL was designated as cqSCN-003.

Seed Protein and Oil

The two major soybean products are seed derived protein meal and oil (Wilson, 2004; Warrington et al., 2015). Soybean cultivars seeds typically average ~40% protein content and ~20% oil content on a dry weight basis. Currently there are 248 and 327 QTL associated with seed protein and seed oil content, respectively, recorded in the Soybean genetics and genomics database (https://www.soybase.org/.org, accessed on 11/12/2020). Many of these QTL were discovered through linkage analysis which requires F_2 generation, backcross, or RIL derived from original biparental crosses (Leamy et al., 2017).

Yaklich et al. (2002) evaluated protein and oil concentration over 51 years (1948 – 1998) using the Northern and Southern Uniform Soybean Tests to determine long-term trends. The Northern and Southern Soybean Tests mean protein concentration was above 420 g kg-1 by 1996. The mean oil concentration was on an upward trend that was greater than 220 g kg^{-1} from 1948 – 1973 but dropped below 200 g kg^{-1} after 1974 and then steadily increased to around 200 g kg^{-1}. Yaklich et al. (2002) suggested that the production environments and unpredictable weather patterns caused the decrease in oil concentration in the both the uniform test trials. The mean protein to oil ratio ranged from 1.82 – 2.10. Andresen et al. (2001) reported climate change with favorable weather conditions for agronomic crops between the mid 1950 – 1970 called the *benign climate* period which could explain the change in oil concentration during the mid-1970s (Yaklich et al., 2002).

A core collection evaluation of *G. max* and *G. soja* seed composition reported 36-40% for *G. max* check lines and 39-48% for *G. soja* accessions for protein concentration and a variation of 21-25% for *G. max* check lines and 15-17% for *G. soja* accessions for oil concentration (La et al., 2019). Based on this and other studies, *G. soja* accessions tend to have more protein content and less oil content than *G. max* check lines. La et al., (2019) reported a negative correlation between protein and oil of -0.66, which was consistent with previous reports (Chung et al., 2013; Leamy et al., 2017).

There are many factors that influence soybean seed protein and seed oil content, such as genetics, environments, genetic and environment interactions, and management practices. Assefa et al. (2019) provided a comprehensive analysis of environments, management practices, and genetics influencing US soybean yield, protein, and oil from 21 studies between 2002 – 2017. Oil content increased at a rate of 1.2 g kg^{-1} per Mg ha^{-1} seed yield increase and protein content

decreased at 1.3 g kg^{-1} per Mg ha^{-1} seed yield increase. Across environments, the proportion of variation explained in mean oil was $R^2 = 0.80$, mean protein was $R^2 = 0.85$, and seed yield was $R^2 = 0.74$. Assefa et al. (2019) concluded that within an environment, seed oil content and seed yield increased from 2002 – 2017, while seed protein content decreased due to the narrow variation in seed oil and seed protein content and the negative correlation between seed oil and seed protein. Management practices such as planting date and fertilizer nitrogen application can affect seed oil content, seed protein content, and seed yield. By delaying planting on a per week basis, seed oil content and seed yield saw a significant decrease rate of -0.007% and -0.011%, respectively (Assefa et al., 2019). Seed protein content was not significantly affected but did decline at a rate of -0.027%. When applying nitrogen fertilizer, less than 50 kg ha^{-1} of nitrogen fertilizer improves seed oil content more than 200 g kg^{-1} and seed protein content more than 400 g $kg^{-}1$, while more than 100 kg ha^{-1} of nitrogen fertilizer will improve seed yield more than 5 Mg ha^{-1} (Assefa et al., 2019).

Diers et al. (1992) discovered two major QTL controlling protein content from a cross between the *G. soja* PI 468916, a high protein wild soybean from Liaoning, China, and the *G. max* line A8-3356022, which is a maturity group III experimental line from Iowa State University. The two reported QTL were on Chr. 15 and 20. The *G. soja* allele for the most significant marker from Chr. 20 and Chr. 15 had an increase in protein of 2.4% and 1.7%, respectively. The QTL on Chr. 20 has been consistently mapped (Diers et al., 1992; Hwang et al., 2014; Warrington et al., 2015) and investigated because of its high additive effect and stability (Lestari et al. 2013). From the Soybean Genetics Committee (https://www.soybase.org/), two QTL, cqSeed protein-001 on Chr. 15 (Nichols et al., 2006) and cqSeed protein-003 on Chr. 20 (Fasoula et al., 2004), have been confirmed.

Diers et al. (1992) first identified a major QTL for seed protein and seed oil content. Since then, many studies have reported QTL regions and molecular markers to be associated with seed protein and seed oil content but most of the QTL have not been confirmed in different backgrounds or applied to breeding programs because of their low phenotypic variation (Van and McHale, 2017). Kim et al. (2016) identified a QTL located on Chr. 15 from a cross of the donor parent, LG00-13329, and the recurrent parent, Williams 82. LG00-13329 derived from a cross between PI 407788A, which is a high seed protein accession from Korea, and Williams 82. The QTL on Chr. 15 was found to be associated with an increased in seed protein content and decreased in seed oil content. Warrington et al. (2015) identified QTL for seed protein and amino acid on Chr. 14, 15, 17, and 20 in the Benning x Danbaekkong population and mapped Chr. 20 which explained 55% of the phenotypic variation and contains the *G. soja* Danbaekkong allele. Sebolt et al. (2000) studied the effect of the Chr. 20 protein QTL (Diers et al., 1992) in three populations in different genetic backgrounds. This led Sebolt et al. (2000) to identify C1914, a high protein line, to have a similar high protein allele on Chr. 20 as *G. soja*. A study from Chung et al. (2003) reported a protein QTL allele mapped to Chr. 20 (Diers et al., 1992) from a *G. max* accession PI 437088A. Chung et al. (2003) protein QTL and Sebolt et al. (2000) protein QTL was mapped to the same region on Chr. 20. Both QTL were associated for increased seed protein content and negatively correlated with seed oil content and yield. This suggest that the *G. soja* Danbaekkong line may have a different allele on Chr. 20 or may have a genetic background that has a lesser effect on yield (Warrington et al., 2015; Patil et al., 2017).

Patil et al., (2018) studied an interspecific mapping population, consisting of 188 $F_{7:8}$ RIL, from a cross between the cultivar Williams 82 and a wild soybean accession PI 483460B. A combination of QTL mapping with BARCSoySNP6K (Song et al., 2020) markers, bin mapping

with skim-whole genome resequencing data, genome-wide association study, and haplotype was used to map novel alleles and QTL for seed composition traits. Patil et al. (2018) identified five QTL for seed protein content on Chr. 6, 8, 13, 19, and 20 and nine QTL for seed oil content on Chr. 2, 7, 8, 9, 14, 15, 17, 19, and 20 by composite interval mapping using bin markers. Two significant protein loci were identified on Chr. 20 and one oil locus was identified on Chr. 5 using GWAS. Haplotype analysis was able to identify multiple QTL and were performed on major significant QTL, qSuc_08 for sucrose content and qPro_20 for protein/oil content. Patil et al. (2018) identified 19 nonsynonymous SNPs underlying qPro_20 where the amino acid change in Glyma.20g096100, Glyma.20g096800, and Glyma.20g097400. Using a large volume of SNP markers derived from the skim-whole genome resequencing data, Patil et al. (2018) was able to develop a very high-resolution bin map and show collinearity of the bin map with the physical map. Narrow QTL interval were identified when comparing bin-based genetic mapping with 3K-SNP which resulted in fewer putative candidate genes that were functionally correlated with traits of interest. Linkage mapping, GWAS, and haplotype analysis was able to support the consistent QTL across all environments and inferred the allelic variation within identified QTL.

Seo et al. (2018) identified 12 QTL for seed protein content, 11 QTL for seed oil content, and 4 QTL for both seed protein and seed oil content using 1570 SNP markers. Among the identified QTL, 6 novel seed protein QTL were found on Chr. 2, 6, 10, 13, 15, and 17, while 4 novel seed oil QTL were found on Chr. 7, 15, 16, and 17. Five seed protein QTL with more than 10% of the phenotypic variance explained for seed protein were identified on Chr. 2, 13, 17, 19, and 20. Four seed oil QTL with more than 10% of the phenotypic variance explained were identified on Chr. 13, 15, 19, and 20. Zhang et al. (2019) used a combination of linkage and GWAS analysis to identified four significant SNP loci regions distributed on Chr. 2, 6, 9, and 20

for seed protein and oil. The linkage analysis panel consisted of RIL from 308 $F_{2:7}$ lines derived from a cross between HH27 and ZGDD. The GWAS panel included 182 accessions from China and 21 accessions from the United States. The two algorithms used in the linkage analysis were inclusive composite interval mapping and a mixed model based on the composite interval mapping. Zhang et al. (2019) reported that the QTL on Chr. 20 had the highest phenotypic variation explained and additive effect with a range of 7.27 – 9.39% and 0.56-0.75, respectively. All the QTL intervals from this study either overlapped or were close to regions reported by previous studies (Diers et al., 1992; Tajuddin et al., 2003; Qi et al., 2011; Pathan et al., 2013; Patil et al., 2018; Seo et al., 2018).

Brzostowki et al. (2017) suggested that it is important to evaluate a QTL across multiple genetic backgrounds and environment because of its complex relationships between seed composition and seed yield before incorporating it into a breeding program. Seed composition and their correlation with each other and with seed yield sis heavily influenced by the environment, genetics, climate, and management practices (Yaklich et al., 2002; Bellaloui et al., 2009; Assefa et al., 2019). The increase in seed protein with a yield drag is not desirable as a suitable cultivar crop. The negative correlation between seed protein and seed oil, and seed protein and yield, makes it challenging to develop a soybean cultivar that has increase seed protein and seed oil content as well as increased yield (Brzostowki et al., 2017).

Fine Mapping

The three broad classes for QTL mapping are regression (Haley and Knott, 1992; Whittaker et al., 1996), maximum likelihood (Doerge et al., 1997), and Bayesian models

(Sillanpaa and Corander 2002). According to (Lander and Botstein, 1989), interval mapping is based on maximum-likelihood parameter estimation and provides a likelihood-ratio test for QTL position. Using regression for interval mapping saves computational time at one or multiple genomic positions and is useful when genetic maps are relatively sparse in markers (Haley and Knott, 1992; Martinez and Curnow 1992). When using interval mapping, the estimates of locations and effects of QTL may be biased when QTL are linked, which is a major disadvantage (Haley and Knott, 1992; Martinez and Curnow 1992; Zeng, 1994). Bayesian models are not widely used in practice due to the difficulty of choosing prior distributions, complexity of computation, and lack of user-friendly software (Li et al., 2006). Composite interval mapping (CIM) is a combination of interval mapping and multiple marker regression analysis (Jansen, 1994; Zeng, 1994). An advantage of CIM is that it controls the effects of QTL on other intervals or chromosomes onto the QTL that is being tested, which allows it to increase the precision of QTL detection.

There are three essential components to fine mapping: (1) SNP in the region need to be genotyped or imputed with high confidence, (2) strict quality control, and (3) a large population size is required to provide enough power to differentiate between SNP in high linkage disequilibrium (LD) (Spain and Barrett, 2015). A strict quality control is needed for accurate imputation to exclude genotyping errors before imputation. This is usually done manually by checking the intensity cluster plots for all associated variants (Spain and Barrett, 2015). Fine-mapping studies are imputed from dense genotyping chip arrays, then an association analysis is performed along with a stepwise conditional analysis (Spain and Barrett, 2015). There are many different methods to identify casual variants to explain marker associations. They can be classified into two groups: triaging variants based on P-values or LD to the lead SNP and

Bayesian methods that assign posterior probabilities of causality to each SNP (Spain and Barrett, 2015).

A single marker regression and haplotype analysis can be useful in determining which markers are associated with a trait of interest, thus reducing the size of a major QTL. Zhang et al. (2012) fine mapped a major flowering time QTL, *qFT6*, on soybean Chr. 6 by combining linkage and association mapping. In this study, 8 SSR markers and 10 SNP markers were used to fine map the QTL region *qFT6* and a total of 11 markers were significantly associated with *qFT6* by general linear model analysis. Zhang et al. (2012) performed haplotype analysis between every two markers and performed regression analysis of the haplotypes to the phenotypic data to determine the joint effect of the loci pairs associated with flowering time. Two marker combination explained the most phenotypic variation and suggested that the candidate gene were in ~300 kbp between the markers. Three genes were identified after a comprehensive analysis (soybean genome information, bioinformatics, and genome comparison), BLASTP queries of the UnifRef database, and synteny analyses between soybean and other dicotyledonous plants.

Whole Genome Resequencing

The advancement of next-generation sequencing (NGS) has become a strong tool in the field of genomics by allowing researchers to sequence whole genomes (Koboldt et al., 2013). Whole genome sequencing can be classified as *de novo* whole genome sequencing (WGS) and whole genome resequencing (WGR) (Fuentes-Pardo and Ruzzante, 2017). WGS consist of the genome being sequence for the first time and WGR compares genomic variation among individuals or population and requires a reference genome for read mapping and variant calling

(Fuentes-Pardo and Ruzzante, 2017). Individual-based WGR obtains high-quality individual genotypes, which requires a high read depth (>30 – 50x depth) to accurately identify SNP, short INDEL, and genotype calling (Nagasaki et al., 2015). Population-based WGR directly obtains population level genomic data (Buerkle and Gompert, 2012). There are many bioinformatic pipelines that includes multiple steps when analyzing whole genome sequencing data. The standard steps are quality control, read mapping to the reference genome, SNP or indel detection, *de novo* genome assembly, genome annotation, phylogenetic tree, and phylogenetic analysis (Oakeson et al., 2017).

Qi et al. (2014) reported 15 major QTL for 11 traits using a combination of WGS data and a high-density-marker WGR linkage map, and identified a novel gene, *GmCHX1*, for salt tolerance in wild soybean. The combination of using WGS and WGR can potentially be an effective methodology for identifying novel genetic information in wild soybean (Qi et al., 2014). Other studies have also leverage the WGR to identify QTL. Pawlowski et al. (2019) identified six QTL using SNP derived from WGR data to be associated with *Rhizophagus intraradices* colonization, which explained 24% of the phenotypic variation. Boudhrious et al. (2019) used 1.5 million SNP from genotyping-by-sequencing and WGR data to discover a new QTL on Chr. 01 associated with resistance to *Sclerotinia sclerotiorum*. Whole genome sequencing data, a high-density-marker QTL map, and functional analysis can help in identifying QTL and genes for agronomic and seed compositional traits to improve our cultivar crops.

Nitrogen Concentration

Soybean seed yield requires large amounts of nitrogen (N) and high photosynthesis rates to reach optimum vegetative growth and reproductive growth because of high seed protein content (Sinclair and De Wit, 1976; Giller and Cadisch, 1995; Ohyama et al., 2017). Soybean can fix atmospheric dinitrogen (N_2) within root nodules through symbiotic association with soil microorganisms, typically of the genus *Bradyrhizobia* (Bohlool et al., 1992). There are three sources by which soybean plants can assimilate N: 1) symbiotic biological N_2 fixation by root nodules (biological nitrogen fixation, BNF), 2) absorbed from soil mineralized N, and 3) from fertilizer application (Harper, 1974; Harper, 1987, Ohyama et al., 2017). High concentration of mineral or fertilizer N application depresses nodule formation and N_2 fixation activity and induces nodule senescence, which can reduce seed yield (Fujikake et al., 2002; Fujikake et al., 2003; Ohyama et al., 2011).

Roots uptake N in the forms of amino acids, ureides, and other N compounds and are transported to the shoot via the xylem (Rentsch et al., 2007). Plant leaves and roots are the largest sinks during the vegetative stages, and flowers, fruits, and seeds are the dominant sinks during the reproductive stages (Masclaux-Daubresse et al., 2010; Tegeder et al., 2017). Soybean seeds requires a high amount of N because soybean seeds contain a high concentration of protein (Sinclair and De Wit, 1976; Giller and Cadisch, 1995; Ohyama et al., 2017). Amino acids and ureides are unloaded from the phloem into seeds during the reproductive stages (Tegeder et al., 2017). Amino acids are the dominant N forms that are transferred into the apoplast and the imported into the seed embryo where they are partition for seed metabolism and storage protein (Tegeder and Rentsch, 2010; Tegeder et al., 2017).

Silva et al. (2011), and Zhou et al. (2019) reported that nitrogen application at the flowering stage significantly influences soybean reproductive growth and grain yield. Zhou et al. (2019) studied two soybean cultivars with three N treatments: continuous high nitrogen content (CHN), continuous low nitrogen (CLN), and enrich nitrogen supply (ENS) at the R1 stage. After 0 hour of CHN treatment, the shoot dry mass of CHN was greater than CLN. The root/shoot rate of ENS was lower than that of CLN after 12 hours and after 24 hours, there were no significant differences between the two (Zhou et al., 2019). Zhou et al. (2019) concluded that the shoot biomass increased more than the root biomass after the ENS at the R1 growth stage. Soybean demands plentiful nutrition and larger canopy for photosynthesis at the R1 stage to increase shoot biomass to benefit reproductive growth (He, 1982; Silva et al., 2011; Zhou et al., 2019). Li et al. (2018) stated that the carbon requirement for organs decreases as plants grow under a prolonged N limitation, but under high N condition, plants require more carbon for nitrogen metabolism for plant growth. A continuous supply of low-level N throughout the growing season may support plant vigor and photosynthetic activity to increase soybean seed yield and help maintain optimal seed composition (Ohyama et al., 2013).

The trade-off between BNF and indigenous soil N supply has been well documented, where BNF decreases as the contribution from indigenous soil N supply increases (Santachiara et al., 2017; Streeter & Wong, 1988). Cafaro La Menza et al. (2017) and Cafaro La Menza et al. (2019) developed a protocol to assess N limitation to answer the question of whether the combined N supply from BNF and soil N supply are sufficient to meet soybean N requirements in highly productive environments. To better understand seasonal N physiological mechanisms (e.g., BNF, aboveground dry matter (ADM) and N accumulation, leaf area index (LAI), photosynthesis, and N mobilization), Cafaro La Menza et al. (2019) observed differences in seed

yield and seed N concentration between soybean crops growing under different N supply. In this study, Cafaro La Menza et al. (2019) reported that full-N had 51 kg N ha^{-1} total accumulation than in zero-N, which translated to 10% greater ADM at the R7 stage for full-N. The full-N treatment also had 4% greater seed and 7% greater seed number than the zero-N treatment (Cafaro La Menza et al., 2019). ADM and N accumulation became different before the R1 stage in the full-N treatment, but both the full-N and zero-N treatments coincided in the post-peak downward slopes after the R5 stage (Cafaro La Menza et al., 2019). Soil N supply accounted for 65% of N accumulation between the VE and R5 stages, while BNF supplied 90% of N during seed filling (Cafaro La Menza et al., 2019). Cafaro La Menza et al. (2019) reported that accumulated ADM between R3 and R6 was greater in the full-N than in the zero-N which led to full-N having higher seed number and the LAI max was reached sooner in the full-N treatment. Mobilization of N between the two treatments were associated with greater accumulation of N at R5 and the full-N treatment was 17% greater when compared to the zero-N treatment (Cafaro La Menza et al., 2019). Cafaro La Menza et al. (2019) concluded that the amount of N mobilized were greater in the full-N treatment vs the zero-N treatment due to the higher accumulation of N in the non-seed ADM at R5. These studies show that plants still rely on the ability to uptake N during the reproductive stages and on the amount of N that are transferred from leaves, roots, or stem to seed sinks for seed development (Masclaux-Daubresse et al., 2010; Ohyama et al., 2013).

Nitrogen QTL

There are several studies on QTL identification for nitrogen fixation in soybeans and other agronomic crops. More than 70 QTL have been identified for nitrogen fixation across all

20 soybean chromosomes (www.soybase.org, accessed 12/10/2020). Bazzer et al. (2020) studied a population of 196 F_6-derived RIL and identified 13 total QTL associated with isotope ^{15}N and with a phenotypic variance ranging from 1.63%-14.39%. These 13 QTL will help us have a better understanding of the genetic basis of nitrogen fixation and the biological regulation of nitrogen fixation. The RIL population was developed from a cross between PI 416997 and PI 567201D. The parents were selected because of their extreme ratio between the isotopes ^{13}C (carbon) and ^{12}C (Bazzer et al., 2020). Isotope ^{15}N best linear unbiased prediction (BLUP) values were used in the QTL mapping because the BLUP values reduces the environmental effect, which increases the accuracy of detection of QTL.

Zhou et al. (2017) identified seven QTL for nitrogen uptake six QTL associated with nitrogen use efficiency eight QTL were detected for plant biomass yield, and two QTL were identified for grain yield in rice. Nitrogen uptake and grain yield were significantly correlated with an R^2 of 0.74 while nitrogen use efficiency and grain yield were not significantly correlated, concluding that increasing nitrogen uptake could potentially improve grain yield (Zhou et al., 2017). Dhanapal et al. (2015) identified 17 SNP associated with nitrogen derived from the atmosphere (NDFA), 19 SNP associated with nitrogen concentration and, 24 SNP associated with carbon/nitrogen (C/N) ratio in soybeans from a GWAS using BLUP values from each environment and the BLUP mean across all environments. The identified SNP were able to indicate 12, 11, and 17 putative loci to be associated with NDFA, nitrogen concentration, and the ration of C/N, respectively (Dhanapal et al., 2015). Nitrogen plays an important role not only in soybeans, but in other crops as well. Understanding how crops partition nitrogen and how they use efficiently use nitrogen will help improve agronomic and seed compositional traits.

References

Abdelghany AM, Zhang S, Azam M, et al (2019) Natural Variation in Fatty Acid Composition of Diverse World Soybean Germplasms Grown in China. Agronomy 10:24. doi: 10.3390/agronomy10010024

Akond M, Liu S, Boney M, et al (2014) Identification of Quantitative Trait Loci (QTL) Underlying Protein, Oil, and Five Major Fatty Acids' Contents in Soybean. American Journal of Plant Sciences 05:158–167. doi: 10.4236/ajps.2014.51021

Akond M, Liu S, Schoener L, et al (2017) A SNP-Based Genetic Linkage Map of Soybean Using the SoySNP6K Illumina Infinium BeadChip Genotyping Array. Plant Genetics, Genomics, and Biotechnology 1:80–89. doi: 10.5147/pggb.v1i3.154

Akpertey A, Belaffif M, Graef GL, et al (2014) Effects of Selective Genetic Introgression from Wild Soybean to Soybean. Crop Science 54:2683–2695. doi: 10.2135/cropsci2014.03.0189

Alonso-Blanco C, Koornneef M, Stam P (1998) The Use of Recombinant Inbred Lines (RIL) for Genetic Mapping. Arabidopsis Protocols 137–146. doi: 10.1385/0-89603-391-0:137

Andresen JA, Alagarswamy G, Rotz CA, et al (2001) Weather Impacts on Maize, Soybean, and Alfalfa Production in the Great Lakes Region, 1895-1996. Agronomy Journal 93:1059–1070. doi: 10.2134/agronj2001.9351059x

Assefa Y, Purcell LC, Salmeron M, et al (2019) Assessing Variation in US Soybean Seed Composition (Protein and Oil). Frontiers in Plant Science. doi: 10.3389/fpls.2019.00298

Bandillo N, Jarquin D, Song Q, et al (2015) A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. The Plant Genome. doi: 10.3835/plantgenome2015.04.0024

Bazzer SK, Kaler AS, Ray JD, et al (2020) Identification of quantitative trait loci for carbon isotope ratio (δ13C) in a recombinant inbred population of soybean. Theoretical and Applied Genetics 133:2141–2155. doi: 10.1007/s00122-020-03586-0

Beche E, Gillman JD, Song Q, et al (2020) Nested association mapping of important agronomic traits in three interspecific soybean populations. Theoretical and Applied Genetics 133:1039–1054. doi: 10.1007/s00122-019-03529-4

Bellaloui N, Smith JR, Ray JD, Gillen AM (2009) Effect of Maturity on Seed Composition in the Early Soybean Production System as Measured on Near-Isogenic Soybean Lines. Crop Science 49:608–620. doi: 10.2135/cropsci2008.04.0192

Boehm JD, Abdel-Haleem H, Schapaugh WT, et al (2019) Genetic Improvement of US Soybean in Maturity Groups V, VI, and VII. Crop Science 59:1838–1852. doi: 10.2135/cropsci2018.10.0627

Boerma HR, Specht JE (2004) Soybeans: improvement, production, and uses. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America

Bohlool BB, Ladha JK, Garrity DP, George T (1992) Biological nitrogen fixation for sustainable agriculture: A perspective. Plant and Soil 141:1–11. doi: 10.1007/bf00011307

Borevitz JO, Nordborg M (2003) The Impact of Genomics on the Study of Natural Variation in Arabidopsis: Figure 1. Plant Physiology 132:718–725. doi: 10.1104/pp.103.023549

Boudhrioua C, Bastien M, Torkamaneh D, Belzile F (2019) Genome-wide association mapping of Sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. doi: 10.21203/rs.2.14709/v2

Boudhrioua C, Bastien M, Torkamaneh D, Belzile F (2019) Genome-wide association mapping of Sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. doi: 10.21203/rs.2.14709/v2

Brechenmacher L, Nguyen TH, Zhang N, et al (2015) Identification of Soybean Proteins and Genes Differentially Regulated in Near Isogenic Lines Differing in Resistance to Aphid Infestation. Journal of Proteome Research 14:4137–4146. doi: 10.1021/acs.jproteome.5b00146

Broman KW (2004) The Genomes of Recombinant Inbred Lines. Genetics 169:1133–1146. doi: 10.1534/genetics.104.035212

Broman KW (2011) Guide to qtl mapping with r/qtl. Springer-Verlag New York

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890. doi: 10.1093/bioinformatics/btg112

Brummer EC, Graef GL, Orf J, et al (1997) Mapping QTL for Seed Protein and Oil Content in Eight Soybean Populations. Crop Science 37:370–378. doi: 10.2135/cropsci1997.0011183x003700020011x

Brzostowski LF, Diers BW (2017) Agronomic Evaluation of a High Protein Allele from PI407788A on Chromosome 15 across Two Soybean Backgrounds. Crop Science 57:2972–2978. doi: 10.2135/cropsci2017.02.0083

Brzostowski LF, Pruski TI, Specht JE, Diers BW (2017) Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. Theoretical and Applied Genetics 130:2315–2326. doi: 10.1007/s00122-017-2961-x

Buerkle A, Gompert Z (2012) Population genomics based on low coverage sequencing: how low should we go? Molecular Ecology 22:3028–3035. doi: 10.1111/mec.12105

Cafaro La Menza N, Monzon JP, Specht JE, et al (2019) Nitrogen limitation in high-yield soybean: Seed yield, N accumulation, and N-use efficiency. Field Crops Research 237:74–81. doi: 10.1016/j.fcr.2019.04.009

Cafaro La Menza N, Monzon JP, Specht JE, Grassini P (2017) Is soybean yield limited by nitrogen supply? Field Crops Research 213:204–212. doi: 10.1016/j.fcr.2017.08.009

Carter TE, Nelson RL, Sneller CH, Cui Z (2016) Genetic Diversity in Soybean. Agronomy Monographs 303–416. doi: 10.2134/agronmonogr16.3ed.c8

Chen Q-shan, Zhang Z-chen, Liu C-yan, et al (2007) QTL Analysis of Major Agronomic Traits in Soybean. Agricultural Sciences in China 6:399–405. doi: 10.1016/s1671-2927(07)60062-5

Clayton, C. (2019, August 12). Farm Service Agency Cites Record Number of Prevented Planting Acres. Retrieved December 28, 2019, from DTN Progressive Farmer website: https://www.dtnpf.com/agriculture/web/ag/news/article/2019/08/12/farm-service-agency-cites-record

Chung J, Babka HL, Graef GL, 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

Collard BC, Jahufer MZ, Brouwer JB, Pang EC (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169–196. doi: 10.1007/s10681-005-1681-5

Concibido V, La Vallee B, Mclaird P, et al (2003) Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. Theoretical and Applied Genetics 106:575–582. doi: 10.1007/s00122-002-1071-5

Cordeiro CF, Echer FR (2019) Interactive Effects of Nitrogen-Fixing Bacteria Inoculation and Nitrogen Fertilization on Soybean Yield in Unfavorable Edaphoclimatic Environments. Scientific Reports. doi: 10.1038/s41598-019-52131-7

Dei HK (2011) Soybean as a Feed Ingredient for Livestock and Poultry. Recent Trends for Enhancing the Diversity and Quality of Soybean Products. doi: 10.5772/17601

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm.

Dhanapal AP, Ray JD, Singh SK, et al (2015) Genome-Wide Association Analysis of Diverse Soybean Genotypes Reveals Novel Markers for Nitrogen Traits. The Plant Genome. doi: 10.3835/plantgenome2014.11.0086

Diers BW, Keim P, Fehr WR, Shoemaker RC (1992) RFLP analysis of soybean seed protein and oil content. Theoretical and Applied Genetics 83:608–612. doi: 10.1007/bf00226905

Doerge RW, Zeng Z-B, Weir BS (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statistical Science 12:195–219. doi: 10.1214/ss/1030037909

Doyle, Jj & Doyle, Jl. (1986). A Rapid DNA Isolation Procedure from Small Quantities of Fresh Leaf Tissues. Phytochem Bull. 19.

Eskandari M, Cober ER, Rajcan I (2013) Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theoretical and Applied Genetics 126:1677–1687. doi: 10.1007/s00122-013-2083-z

Evangelista JS, Alves RS, Peixoto MA, et al (2021) Soybean productivity, stability, and adaptability through mixed model methodology. Ciência Rural. doi: 10.1590/0103-8478cr20200406

Fabre F, Planchon C (2000) Nitrogen nutrition, yield and protein content in soybean. Plant Science 152:51–58. doi: 10.1016/s0168-9452(99)00221-6

Falke KC, Frisch M (2010) Power and false-positive rate in QTL detection with near-isogenic line libraries. Heredity 106:576–584. doi: 10.1038/hdy.2010.87

Fasoula VA, Harris DK, Boerma HR (2004) Validation and Designation of Quantitative Trait Loci for Seed Protein, Seed Oil, and Seed Weight from Two Soybean Populations. Crop Science 44:1218–1225. doi: 10.2135/cropsci2004.1218

FAOSTAT, www.fao.org/faostat/en/#search/soybean. Accessed 3/01/2021.

Fox CM, Cary TR, Colgrove AL, et al (2013) Estimating Soybean Genetic Gain for Yield in the Northern United States-Influence of Cropping History. Crop Science 53:2473–2482. doi: 10.2135/cropsci2012.12.0687

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proceedings of the National Academy of Sciences 97:4718–4723. doi: 10.1073/pnas.97.9.4718

Fritschi FB, Ray JD, Purcell LC, et al (2013) DIVERSITY AND IMPLICATIONS OF SOYBEAN STEM NITROGEN CONCENTRATION. Journal of Plant Nutrition 36:2111–2131. doi: 10.1080/01904167.2012.748800

Fu Y-B (2015) Understanding crop genetic diversity under modern plant breeding. Theoretical and Applied Genetics 128:2131–2142. doi: 10.1007/s00122-015-2585-y

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology 26:5369–5406. doi: 10.1111/mec.14264

Fujikake H (2003) Quick and reversible inhibition of soybean root nodule growth by nitrate involves a decrease in sucrose supply to nodules. Journal of Experimental Botany 54:1379–1388. doi: 10.1093/jxb/erg147

Fujikake H, Yashima H, Sato T, et al (2002) Rapid and reversible nitrate inhibition of nodule growth and N2fixation activity in soybean (Glycine max(L.) Merr.). Soil Science and Plant Nutrition 48:211–217. doi: 10.1080/00380768.2002.10409193

Funatsuki H, Suzuki M, Hirose A, et al (2014) Molecular basis of a shattering resistance boosting global dissemination of soybean. Proceedings of the National Academy of Sciences 111:17797–17802. doi: 10.1073/pnas.1417282111

Furuta T, Ashikari M, Jena KK, et al (2017) Adapting Genotyping-by-Sequencing for Rice F2 Populations. G3: Genes|Genomes|Genetics 7:881–893. doi: 10.1534/g3.116.038190

Geladi P, MacDougall D, Martens H (1985) Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Applied Spectroscopy 39:491–500. doi: 10.1366/0003702854248656

Gelli M, Mitchell SE, Liu K, et al (2016) Mapping QTLs and association of differentially expressed gene transcripts for multiple agronomic traits under different nitrogen levels in sorghum. BMC Plant Biology. doi: 10.1186/s12870-015-0696-x

Giller KE, Cadisch G (1995) Future benefits from biological nitrogen fixation: An ecological approach to agriculture. Management of Biological Nitrogen Fixation for the Development of More Productive and Sustainable Agricultural Systems 255–277. doi: 10.1007/978-94-011-0053-3_13

Gillman JD, Tetlow A, Lee J-D, et al (2011) Loss-of-function mutations affecting a specific Glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. BMC Plant Biology 11:155. doi: 10.1186/1471-2229-11-155

Gizlice Z, Carter TE, Burton JW (1993) Genetic Diversity in North American Soybean: I. Multivariate Analysis of Founding Stock and Relation to Coefficient of Parentage. Crop Science 33:614–620. doi: 10.2135/cropsci1993.0011183x003300030038x

Glover KD, Wang D, Arelli PR, et al (2004) Near Isogenic Lines Confirm a Soybean Cyst Nematode Resistance Gene from PI 88788 on Linkage Group J. Crop Science 44:1505–1505. doi: 10.2135/cropsci2004.1505a

Goodstein DM, Shu S, Howson R, et al (2011) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research. doi: 10.1093/nar/gkr944

Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. Trends in Plant Science 15:529–537. doi: 10.1016/j.tplants.2010.05.008

Guo B, Sleper DA, Arelli PR, et al (2006) Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763. Theoretical and Applied Genetics 112:984–985. doi: 10.1007/s00122-005-0150-9

Guo J, Wang Y, Song C, et al (2010) A single origin and moderate bottleneck during domestication of soybean (Glycine max): implications from microsatellites and nucleotide sequences. Annals of Botany 106:505–514. doi: 10.1093/aob/mcq125

Ha B-K, Vuong TD, Velusamy V, et al (2013) Genetic mapping of quantitative trait loci conditioning salt tolerance in wild soybean (Glycine soja) PI 483463. Euphytica 193:79–88. doi: 10.1007/s10681-013-0944-9

Haldane JB, Waddington CH (1931) INBREEDING AND LINKAGE. Genetics 16:504–504. doi: 10.1093/genetics/16.5.504a

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324. doi: 10.1038/hdy.1992.131

Hammond EG, Johnson LA, Su C, et al (2005) Soybean Oil. Bailey's Industrial Oil and Fat Products. doi: 10.1002/047167849x.bio041

Harlan JR, de Wet JM, Price EG (1973) Comparative Evolution of Cereals. Evolution 27:311. doi: 10.2307/2406971

Harper JE (1974) Soil and Symbiotic Nitrogen Requirements for Optimum Soybean Production 1. Crop Science 14:255–260. doi: 10.2135/cropsci1974.0011183x001400020026x

Harper JE, Nicholas JC (1978) Nitrogen Metabolism of Soybeans. Plant Physiology 62:662–664. doi: 10.1104/pp.62.4.662

Hartwig EE, Kilen TC (1991) Yield and Composition of Soybean Seed from Parents with Different Protein, Similar Yield. Crop Science 31:290–292. doi: 10.2135/cropsci1991.0011183x003100020011x

Heim CB, Gillman JD (2016) Genotyping-by-Sequencing-Based Investigation of the Genetic Architecture Responsible for a ~Sevenfold Increase in Soybean Seed Stearic Acid. G3: Genes|Genomes|Genetics 7:299–308. doi: 10.1534/g3.116.035741

Huang J, Ma Q, Cai Z, et al (2020) Identification and Mapping of Stable QTLs for Seed Oil and Protein Content in Soybean [Glycine max(L.) Merr.]. Journal of Agricultural and Food Chemistry 68:6448–6460. doi: 10.1021/acs.jafc.0c01271

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

Hymowitz T, Collins FI, Panczner J, Walker WM (1972) Relationship Between the Content of Oil, Protein, and Sugar in Soybean Seed 1. Agronomy Journal 64:613–616. doi: 10.2134/agronj1972.00021962006400050019x

Hyten DL, Pantalone VR, Sams CE, et al (2004) Seed quality QTL in a prominent soybean population. Theoretical and Applied Genetics 109:552–561. doi: 10.1007/s00122-004-1661-5

Hyten DL, Song Q, Zhu Y, et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences 103:16666–16671. doi: 10.1073/pnas.0604379103

Jaganathan D, Bohra A, Thudi M, Varshney RK (2020) Fine mapping and gene cloning in the post-NGS era: advances and prospects. Theoretical and Applied Genetics 133:1791–1810. doi: 10.1007/s00122-020-03560-w

Jander G, Norris SR, Rounsley SD, et al (2002) Arabidopsis Map-Based Cloning in the Post-Genome Era. Plant Physiology 129:440–450. doi: 10.1104/pp.003533

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455. doi: 10.1093/genetics/136.4.1447

Johnson HW, Robinson HF, Comstock RE (1955) Genotypic and Phenotypic Correlations in Soybeans and Their Implications in Selection 1. Agronomy Journal 47:477–483. doi: 10.2134/agronj1955.00021962004700100008x

Kabelka EA, Diers BW, Fehr WR, et al (2004) Putative Alleles for Increased Yield from Soybean Plant Introductions. Crop Science 44:784–791. doi: 10.2135/cropsci2004.7840

Kadam S, Vuong TD, Qiu D, et al (2016) Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. Plant Science 242:342–350. doi: 10.1016/j.plantsci.2015.08.015

Kakiuchi J, Kobata T (2008) High Carbon Requirements for Seed Production in Soybeans [Glycine max(L.) Merr.]. Plant Production Science 11:198–202. doi: 10.1626/pps.11.198

Kaur G, Serson W, Orlowski J, et al (2017) Nitrogen Sources and Rates Affect Soybean Seed Composition in Mississippi. Agronomy 7:77. doi: 10.3390/agronomy7040077

Keurentjes JJ, Bentsink L, Alonso-Blanco C, et al (2006) Development of a Near-Isogenic Line Population ofArabidopsis thalianaand Comparison of Mapping Power With a Recombinant Inbred Line Population. Genetics 175:891–905. doi: 10.1534/genetics.106.066423

Kim JH, Bae DN, Park S-K, et al (2017) Molecular Genetic Analysis of a Novel Recessive White Flower Gene in Wild Soybean. Crop Science 57:3027–3034. doi: 10.2135/cropsci2017.03.0163

Kim M, Hyten DL, Bent AF, Diers BW (2010) Fine Mapping of the SCN Resistance Locus rhg1-b from PI 88788. The Plant Genome. doi: 10.3835/plantgenome2010.02.0001

Kim M, Schultz S, Nelson RL, Diers BW (2016) Identification and Fine Mapping of a Soybean Seed Protein QTL from PI 407788A on Chromosome 15. Crop Science 56:219–225. doi: 10.2135/cropsci2015.06.0340

Koboldt DC, Steinberg KM, Larson DE, et al (2013) The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell 155:27–38. doi: 10.1016/j.cell.2013.09.006

Kole C (2014) Wild Crop Relatives: Genomic and Breeding Resources Legume Crops and Forages. Springer Berlin

La T, Large E, Taliercio E, 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

La TC, Scaboo A (2018) Characterization of a diverse USDA collection of wild soybean (glycine soja siebold & zucc.) accessions and subsequent mapping for seed composition and agronomic traits in a RIL population. (Doctoral disstertation) Retrieved from https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/66386/research.pdf?sequence=1&isAllowed=y

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199. doi: 10.1093/genetics/121.1.185

Leamy LJ, Zhang H, Li C, et al (2017) A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC Genomics. doi: 10.1186/s12864-016-3397-4

Lee J, Hwang Y-S, Kim ST, et al (2017) Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. Food Chemistry 214:248–258. doi: 10.1016/j.foodchem.2016.07.066

Lee J-D, Yu J-K, Hwang Y-H, et al (2008) Genetic Diversity of Wild Soybean (Glycine soja Sieb. and Zucc.) Accessions from South Korea and Other Countries. Crop Science 48:606–616. doi: 10.2135/cropsci2007.05.0257

Lee SH, Bailey MA, Mian MA, et al (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theoretical and Applied Genetics 93-93:649–657. doi: 10.1007/bf00224058

Leffel RC, Cregan PB, Bolgiano AP, Thibeau DJ (1992) Nitrogen Metabolism of Normal and High-Seed-Protein Soybean. Crop Science 32:747–750. doi: 10.2135/cropsci1992.0011183x003200030034x

Lestari P, Van K, Lee J, et al (2013) Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. Frontiers in Plant Science. doi: 10.3389/fpls.2013.00176

Li D, Pfeiffer TW, Cornelius PL (2008) Soybean QTL for Yield and Yield Components Associated with Glycine soja Alleles. Crop Science 48:571–581. doi: 10.2135/cropsci2007.06.0361

Li H, Ye G, Wang J (2006) A Modified Algorithm for the Improvement of Composite Interval Mapping. Genetics 175:361–374. doi: 10.1534/genetics.106.066811

Li M-W, Muñoz NB, Wong C-F, et al (2016) QTLs Regulating the Contents of Antioxidants, Phenolics, and Flavonoids in Soybean Seeds Share a Common Genomic Region. Frontiers in Plant Science. doi: 10.3389/fpls.2016.00854

Li Y, Guan R, Liu Z, et al (2008) Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. Theoretical and Applied Genetics 117:857–871. doi: 10.1007/s00122-008-0825-0

Li Y, Yu Z, Jin J, et al (2018) Impact of Elevated CO2 on Seed Quality of Soybean at the Fresh Edible and Mature Stages. Frontiers in Plant Science. doi: 10.3389/fpls.2018.01413

Li Y-hui, Zhou G, Ma J, et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology 32:1045–1052. doi: 10.1038/nbt.2979

Li, Xu, Yang, Zhao (2019) Dissecting the Genetic Architecture of Seed Protein and Oil Content in Soybean from the Yangtze and Huaihe River Valleys Using Multi-Locus Genome-Wide Association Studies. International Journal of Molecular Sciences 20:3041. doi: 10.3390/ijms20123041

Liu B, Fujita T, Yan Z-H, et al (2007) QTL Mapping of Domestication-related Traits in Soybean (Glycine max). Annals of Botany 100:1027–1038. doi: 10.1093/aob/mcm149

Liu S, Kandoth PK, Lakhssassi N, et al (2017) The soybean GmSNAP18 gene underlies two types of resistance to soybean cyst nematode. Nature Communications. doi: 10.1038/ncomms14822

Liu Z, Li H, Fan X, et al (2016) Phenotypic Characterization and Genetic Dissection of Growth Period Traits in Soybean (Glycine max) Using Association Mapping. PLOS ONE. doi: 10.1371/journal.pone.0158602

Lu W, Wen Z, Li H, et al (2012) Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. Theoretical and Applied Genetics 126:425–433. doi: 10.1007/s00122-012-1990-8

Martínez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoretical and Applied Genetics 85:480–488. doi: 10.1007/bf00222330

Masclaux-Daubresse C, Daniel-Vedele F, Dechorgnat J, et al (2010) Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. Annals of Botany 105:1141–1157. doi: 10.1093/aob/mcq028

Masuda, Tadayoshi & Goldsmith, Peter. (2009). World Soybean Production: Area Harvested, Yield, and Long-Term Projections. International Food and Agribusiness Management Review. 12.

McKenna A, Hanna M, Banks E, 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

Medic J, Atkinson C, Hurburgh CR (2014) Current Knowledge in Soybean Composition. Journal of the American Oil Chemists' Society 91:363–384. doi: 10.1007/s11746-013-2407-9

Mello Filho OL, Sediyama CS, Moreira MA, et al (2004) Grain yield and seed quality of soybean selected for high protein content. Pesquisa Agropecuária Brasileira 39:445–450. doi: 10.1590/s0100-204x2004000500006

Merry R, Butenhoff K, Campbell BW, et al (2019) Identification and Fine-Mapping of a Soybean Quantitative Trait Locus on Chromosome 5 Conferring Tolerance to Iron Deficiency Chlorosis. The Plant Genome 12:190007. doi: 10.3835/plantgenome2019.01.0007

Miao L, Yang S, Zhang K, et al (2019) Natural variation and selection in GmSWEET39 affect soybean seed oil content. New Phytologist 225:1651–1666. doi: 10.1111/nph.16250

Morr CV (1981) Nitrogen Conversion Factors for Several Soybean Protein Products. Journal of Food Science 46:1362–1363. doi: 10.1111/j.1365-2621.1981.tb04175.x

Muehlbauer GJ, Specht JE, Thomas-Compton MA, et al (1988) Near-Isogenic Lines—A Potential Resource in the Integration of Conventional and Molecular Marker Linkage Maps. Crop Science 28:729–735. doi: 10.2135/cropsci1988.0011183x002800050002x

Murphy PA, Resurreccion AP (1984) Varietal and environmental differences in soybean glycinin and .beta.-conglycinin content. Journal of Agricultural and Food Chemistry 32:911–915. doi: 10.1021/jf00124a052

Naegle E, Kwanyuen P, Burton J, et al (2008) Seed Nitrogen Mobilization in Soybean: Effects of Seed Nitrogen Content and External Nitrogen Fertility. Journal of Plant Nutrition 31:367–379. doi: 10.1080/01904160801894921

Nagasaki M, Yasuda J, Katsuoka F, et al (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nature Communications. doi: 10.1038/ncomms9018

Nascimento D, Polo LR, Lazzari F, et al (2018) Genomic Association between SNP Markers and QTLs for Protein and Oil Content in Grain Weight in Soybean (Glycine max). Journal of Scientific Research and Reports 20:1–13. doi: 10.9734/jsrr/2018/44150

Nawaz MA, Yang SH, Chung G (2018) Wild Soybeans: An Opportunistic Resource for Soybean Improvement. Rediscovery of Landraces as a Resource for the Future. doi: 10.5772/intechopen.74973

Nichols DM, Glover KD, Carlson SR, et al (2006) Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits. Crop Science 46:834–839. doi: 10.2135/cropsci2005.05-0168

Oakeson KF, Wagner JM, Mendenhall M, et al (2017) Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerging Infectious Diseases 23:1441–1445. doi: 10.3201/eid2309.170416

Ohyama T, Minagawa R, Ishikawa S, et al (2013) Soybean Seed Production and Nitrogen Nutrition. A Comprehensive Survey of International Soybean Research - Genetics, Physiology, Agronomy and Nitrogen Relationships. doi: 10.5772/52287

Ohyama T, Tewari K, Ishikawa S, et al (2017) Role of Nitrogen on Growth and Seed Yield of Soybean and a New Fertilization Technique to Promote Nitrogen Fixation and Seed Yield. Soybean - The Basis of Yield, Biomass and Productivity. doi: 10.5772/66743

Pantalone VR, Rebetzke GJ, Burton JW, Wilson RF (1997) Genetic regulation of linolenic acid concentration in wild soybean Glycine soja accessions. Journal of the American Oil Chemists' Society 74:159–163. doi: 10.1007/s11746-997-0162-5

Park ST, Kim J (2016) Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. International Neurourology Journal. doi: 10.5213/inj.1632742.371

Pathan SM, Vuong T, Clark K, et al (2013) Genetic Mapping and Confirmation of Quantitative Trait Loci for Seed Protein and Oil Contents and Seed Weight in Soybean. Crop Science 53:765–774. doi: 10.2135/cropsci2012.03.0153

Patil G, Chaudhary J, Vuong TD, et al (2017) Development of SNP Genotyping Assays for Seed Composition Traits in Soybean. International Journal of Plant Genomics 2017:1–12. doi: 10.1155/2017/6572969

Patil G, Do T, Vuong TD, et al (2016) Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. Scientific Reports. doi: 10.1038/srep19199

Patil G, Mian R, Vuong T, 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 TD, Kale S, 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

Pawlowski ML, Vuong TD, Valliyodan B, et al (2019) Whole-genome resequencing identifies quantitative trait loci associated with mycorrhizal colonization of soybean. Theoretical and Applied Genetics 133:409–417. doi: 10.1007/s00122-019-03471-5

Peiffer, Gregory A., et al. "Identification of Candidate Genes Underlying an Iron Efficiency Quantitative Trait Locus in Soybean." *Plant Physiology*, vol. 158, no. 4, 2012, pp. 1745–1754., doi:10.1104/pp.111.189860.

Pollard DA (2012) Design and Construction of Recombinant Inbred Lines. Methods in Molecular Biology 31–39. doi: 10.1007/978-1-61779-785-9_3

Pratap A, Das A, Kumar S, Gupta S (2021) Current Perspectives on Introgression Breeding in Food Legumes. Frontiers in Plant Science. doi: 10.3389/fpls.2020.589189

Priolli RH, Carvalho CR, Bajay MM, et al (2019) Genome analysis to identify SNPs associated with oil content and fatty acid components in soybean. Euphytica. doi: 10.1007/s10681-019-2378-5

Qi X, Li M-W, Xie M, et al (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nature Communications. doi: 10.1038/ncomms5340

Qi Z-ming, Wu Q, Han X, et al (2011) Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179:499–514. doi: 10.1007/s10681-011-0386-1

Rainbird RM, Thorne JH, Hardy RW (1984) Role of Amides, Amino Acids, and Ureides in the Nutrition of Developing Soybean Seeds. Plant Physiology 74:329–334. doi: 10.1104/pp.74.2.329

Ray JD, Dhanapal AP, Singh SK, et al (2015) Genome-Wide Association Study of Ureide Concentration in Diverse Maturity Group IV Soybean [Glycine max (L.) Merr.] Accessions. G3: Genes|Genomes|Genetics 5:2391–2403. doi: 10.1534/g3.115.021774

Ray JD, Fritschi FB, Heatherly LG (2006) Large applications of fertilizer N at planting affects seed protein and oil concentration and yield in the Early Soybean Production System. Field Crops Research 99:67–74. doi: 10.1016/j.fcr.2006.03.006

Ray JD, Heatherly LG, Fritschi FB (2006) Influence of Large Amounts of Nitrogen on NoNIRSrigated and Irrigated Soybean. Crop Science 46:52–60. doi: 10.2135/cropsci2005.0043

Rentsch D, Schmidt S, Tegeder M (2007) Transporters for uptake and allocation of organic nitrogen compounds in plants. FEBS Letters 581:2281–2289. doi: 10.1016/j.febslet.2007.04.013

Rincker K, Nelson R, Specht J, et al (2014) Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. Crop Science 54:1419–1432. doi: 10.2135/cropsci2013.10.0665

Roach DA, Wulff RD (1987) Maternal Effects in Plants. Annual Review of Ecology and Systematics 18:209–235. doi: 10.1146/annurev.es.18.110187.001233

Salvagiotti F, Cassman KG, Specht JE, et al (2008) Nitrogen uptake, fixation and response to fertilizer N in soybeans: A review. Field Crops Research 108:1–13. doi: 10.1016/j.fcr.2008.03.001

Santachiara G, Borrás L, Salvagiotti F, et al (2017) Relative importance of biological nitrogen fixation and mineral uptake in high yielding soybean cultivars. Plant and Soil 418:191–203. doi: 10.1007/s11104-017-3279-9

Santos MA, Geraldi IO, Garcia AA, et al (2013) Mapping of QTLs associated with biological nitrogen fixation traits in soybean. Hereditas 150:17–25. doi: 10.1111/j.1601-5223.2013.02275.x

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19:491–504. doi: 10.1038/s41576-018-0016-z

Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi: 10.1038/nature08670

Sebolt AM, Shoemaker RC, Diers BW (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

Sedivy EJ, Wu F, Hanzawa Y (2017) Soybean domestication: the origin, genetic architecture and molecular bases. New Phytologist 214:539–553. doi: 10.1111/nph.14418

Seo J-H, Kim K-S, Ko J-M, et al (2018) Quantitative trait locus analysis for soybean (Glycine max) seed protein and oil concentrations using selected breeding populations. Plant Breeding 138:95–104. doi: 10.1111/pbr.12659

Sillanpää MJ, Arjas E (1998) Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data. Genetics 148:1373–1388. doi: 10.1093/genetics/148.3.1373

Silva MA, Muniz AS, Mannigel AR, et al (2011) Monitoring and evaluation of need for nitrogen fertilizer topdressing for maize leaf chlorophyll readings and the relationship with grain yield. Brazilian Archives of Biology and Technology 54:665–674. doi: 10.1590/s1516-89132011000400004

Sinclair TR, de Wit CT (1975) Photosynthate and Nitrogen Requirements for Seed Production by Various Crops. Science 189:565–567. doi: 10.1126/science.189.4202.565

Skoneczka JA, Maroof MA, Shang C, Buss GR (2009) Identification of Candidate Gene Mutation Associated With Low Stachyose Phenotype in Soybean Line PI200508. Crop Science 49:247–255. doi: 10.2135/cropsci2008.07.0403

Song J, Liu Z, Hong H, et al (2016) Identification and Validation of Loci Governing Seed Coat Color by Combining Association Mapping and Bulk Segregation Analysis in Soybean. PLOS ONE. doi: 10.1371/journal.pone.0159064

Song Q, Hyten DL, Jia G, et al (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE. doi: 10.1371/journal.pone.0054985

Song Q, Hyten DL, Jia G, et al (2015) Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3: Genes|Genomes|Genetics 5:1999–2006. doi: 10.1534/g3.115.019000

Song Q, Yan L, Quigley C, et al (2020) Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. The Plant Journal 104:800–811. doi: 10.1111/tpj.14960

Spain SL, Barrett JC (2015) Strategies for fine-mapping complex traits. Human Molecular Genetics. doi: 10.1093/hmg/ddv260

Specht JE, Hume DJ, Kumudini SV (1999) Soybean Yield Potential-A Genetic and Physiological Perspective. Crop Science 39:1560–1570. doi: 10.2135/cropsci1999.3961560x

Spielbauer G, Armstrong P, Baier JW, et al (2009) High-Throughput Near-Infrared Reflectance Spectroscopy for Predicting Quantitative and Qualitative Composition Phenotypes of Individual Maize Kernels. Cereal Chemistry Journal 86:556–564. doi: 10.1094/cchem-86-5-0556

Stein HH, Berger LL, Drackley JK, et al (2008) Nutritional Properties and Feeding Values of Soybeans and Their Coproducts. Soybeans 613–660. doi: 10.1016/b978-1-893997-64-6.50021-4

Steketee CJ, Sinclair TR, Riar MK, et al (2019) Unraveling the genetic architecture for carbon and nitrogen related traits and leaf hydraulic conductance in soybean using genome-wide association analyses. BMC Genomics. doi: 10.1186/s12864-019-6170-7

Streeter J, Wong PP (1988) Inhibition of legume nodule formation and N2fixation by nitrate. Critical Reviews in Plant Sciences 7:1–23. doi: 10.1080/07352688809382257

Stupar RM (2010) Into the wild: The soybean genome meets its undomesticated relative. Proceedings of the National Academy of Sciences 107:21947–21948. doi: 10.1073/pnas.1016809108

Sundaramoorthy J, Park GT, Chang JH, et al (2016) Identification and Molecular Analysis of Four New Alleles at the W1 Locus Associated with Flower Color in Soybean. PLOS ONE. doi: 10.1371/journal.pone.0159865

USB, 2019. United Soybean Board Supply & Disappearance. USB Market View Database (n.d.). Available at: https://marketviewdb.centrec.com/sd/. (Accessed: March 3, 2021)

Uses for Soybeans | Missouri Soybean. (2012). Retrieved December 27, 2019, from Mosoy.org website: https://mosoy.org/check-off-at-work/domestic-marketing/

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of Quantitative Trait Loci for Protein and Lipid Contents in Soybean Seeds Using Recombinant Inbred Lines. Breeding Science 53:133–140. doi: 10.1270/jsbbs.53.133

Tamagno S, Balboa GR, Assefa Y, et al (2017) Nutrient partitioning and stoichiometry in soybean: A synthesis-analysis. Field Crops Research 200:18–27. doi: 10.1016/j.fcr.2016.09.019

Tamagno S, Sadras VO, Haegele JW, et al (2018) Interplay between nitrogen fertilizer and biological nitrogen fixation in soybean: implications on seed yield and biomass allocation. Scientific Reports. doi: 10.1038/s41598-018-35672-1

Tegeder M, Masclaux-Daubresse C (2017) Source and sink mechanisms of nitrogen transport and use. New Phytologist 217:35–53. doi: 10.1111/nph.14876

Tegeder M, Rentsch D (2010) Uptake and Partitioning of Amino Acids and Peptides. Molecular Plant 3:997–1011. doi: 10.1093/mp/ssq047

Truong Q, Koch K, Yoon JM, et al (2013) Influence of carbon to nitrogen ratios on soybean somatic embryo (cv. Jack) growth and composition. Journal of Experimental Botany 64:2985–2995. doi: 10.1093/jxb/ert138

Uga Y, Sugimoto K, Ogawa S, et al (2013) Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. Nature Genetics 45:1097–1102. doi: 10.1038/ng.2725

USB, 2019. USB Market View Database Legacy, marketviewdb.centrec.com/sd/. Accessed 3/01/2021.

USDA Soybean Germplasm Collection. In: GBIF. https://www.gbif.org/grscicoll/collection/6e5b27ae-183f-47c1-8a60-7dda5fe05b11. Accessed 10/12/2020.

Van K, McHale L (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 18:1180. doi: 10.3390/ijms18061180

von Korff M, Wang H, Léon J, Pillen K (2004) Development of candidate introgression lines using an exotic barley accession (Hordeum vulgare ssp. spontaneum) as donor. Theoretical and Applied Genetics 109:1736–1745. doi: 10.1007/s00122-004-1818-2

Wang J, Chen P, Wang D, et al (2015) Identification and mapping of stable QTL for protein content in soybean seeds. Molecular Breeding. doi: 10.1007/s11032-015-0285-6

Wang K-J, Takahata Y (2007) A Preliminary Comparative Evaluation of Genetic Diversity between Chinese and Japanese Wild Soybean (Glycine soja) Germplasm Pools using SSR markers. Genetic Resources and Crop Evolution 54:157–165. doi: 10.1007/s10722-005-2641-6

Wang P-wu, Di Q, Liu X-Y (2020) Genome-Wide association Study Identifies Candidate Genes Related to Oleic acid content of Soybean Seed. doi: 10.21203/rs.3.rs-17853/v1

Warrington CV, Abdel-Haleem H, Hyten DL, 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

Watanabe S, Xia Z, Hideshima R, et al (2011) A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals that theGIGANTEAGene Is Involved in Soybean Maturity and Flowering. Genetics 188:395–407. doi: 10.1534/genetics.110.125062

Whittaker JC, Thompson R, Visscher PM (1996) On the mapping of QTL by regression of phenotype on marker-type. Heredity 77:23–32. doi: 10.1038/hdy.1996.104

Wilson, R.F. 2004. Seed composition. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, Production, and Uses. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.: 621-677

Wood CW, Torbert HA, Weaver DB (1993) Nitrogen Fertilizer Effects on Soybean Growth, Yield, and Seed Composition. Journal of Production Agriculture 6:354–360. doi: 10.2134/jpa1993.0354

Xia Z, Watanabe S, Yamada T, et al (2012) Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proceedings of the National Academy of Sciences. doi: 10.1073/pnas.1117982109

Xu G, Fan X, Miller AJ (2012) Plant Nitrogen Assimilation and Use Efficiency. Annual Review of Plant Biology 63:153–182. doi: 10.1146/annurev-arplant-042811-105532

Yaklich RW, 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. doi: 10.2135/cropsci2002.1504

Yang Q, Yang Y, Xu R, et al (2019) Genetic Analysis and Mapping of QTLs for Soybean Biological Nitrogen Fixation Traits Under Varied Field Conditions. Frontiers in Plant Science. doi: 10.3389/fpls.2019.00075

Ye H, Song L, Chen H, et al (2018) A major natural genetic variation associated with root system architecture and plasticity improves waterlogging tolerance and yield in soybean. Plant, Cell & Environment. doi: 10.1111/pce.13190

Young ND, Zamir D, Ganal MW, Tanksley SD (1988) Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the tm-2a gene in tomato. Genetics 120:579–585. doi: 10.1093/genetics/120.2.579

Yu X, Yuan F, Fu X, Zhu D (2016) Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. Food Chemistry 196:776–782. doi: 10.1016/j.foodchem.2015.09.092

Yuan G, Wan Y, Li X, et al (2017) Development of Near-Isogenic Lines in a Parthenogenetically Reproduced Thrips Species, Frankliniella occidentalis. Frontiers in Physiology. doi: 10.3389/fphys.2017.00130

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468. doi: 10.1093/genetics/136.4.1457

Zhang D, Cheng H, Hu Z, et al (2012) Fine mapping of a major flowering time QTL on soybean chromosome 6 combining linkage and association analysis. Euphytica 191:23–33. doi: 10.1007/s10681-012-0840-8

Zhang H, Song Q, Griffin JD, Song B-H (2017) Genetic architecture of wild soybean (Glycine soja) response to soybean cyst nematode (Heterodera glycines). Molecular Genetics and Genomics 292:1257–1265. doi: 10.1007/s00438-017-1345-x

Zhang J, Wang X, Lu Y, et al (2018) Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. Molecular Plant 11:460–472. doi: 10.1016/j.molp.2017.12.016

Zhang T, Wu T, Wang L, et al (2019) A Combined Linkage and GWAS Analysis Identifies QTLs Linked to Soybean Seed Protein and Oil Content. International Journal of Molecular Sciences 20:5915. doi: 10.3390/ijms20235915

Zhang W-K, Wang Y-J, Luo G-Z, et al (2004) QTL mapping of ten agronomic traits on the soybean (Glycine max L. Merr.) genetic map and their association with EST markers. Theoretical and Applied Genetics 108:1131–1139. doi: 10.1007/s00122-003-1527-2

Zhang YH, Liu MF, He JB, et al (2015) Marker-assisted breeding for transgressive seed protein content in soybean [Glycine max (L.) Merr.]. Theoretical and Applied Genetics 128:1061–1072. doi: 10.1007/s00122-015-2490-4

Zhong C, Sun S, Zhang X, et al (2020) Fine Mapping, Candidate Gene Identification and Co-segregating Marker Development for the Phytophthora Root Rot Resistance Gene RpsYD25. Frontiers in Genetics. doi: 10.3389/fgene.2020.00799

Zhou H, Yao X, Zhao Q, et al (2019) Rapid Effect of Nitrogen Supply for Soybean at the Beginning Flowering Stage on Biomass and Sucrose Metabolism. Scientific Reports. doi: 10.1038/s41598-019-52043-6

Zhou Y, Tao Y, Tang D, et al (2017) Identification of QTL Associated with Nitrogen Uptake and Nitrogen Use Efficiency Using High Throughput Genotyped CSSLs in Rice (Oryza sativa L.). Frontiers in Plant Science. doi: 10.3389/fpls.2017.01166

Zhou Z, Jiang Y, Wang Z, 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

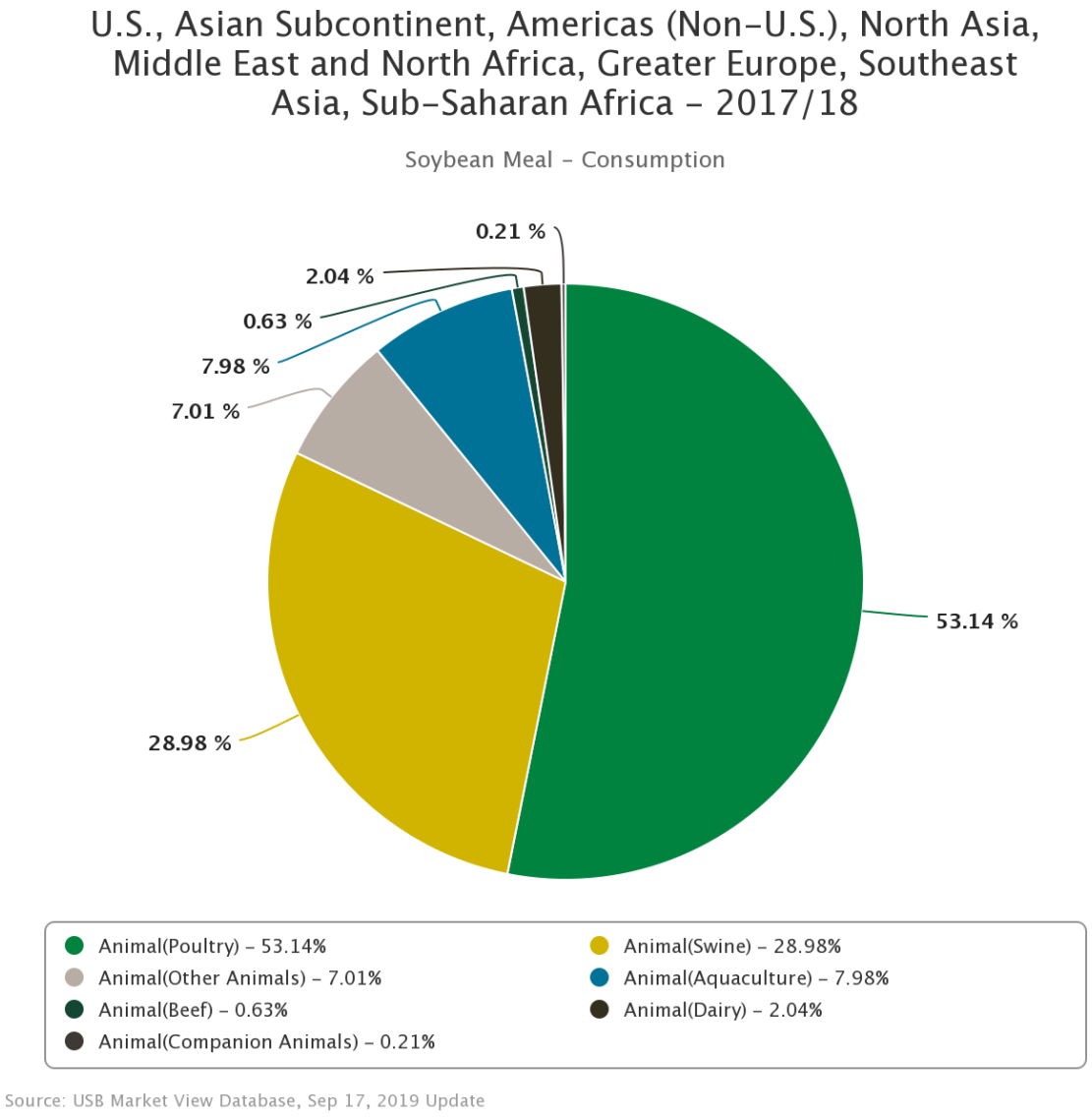

Figure 1-1. Percentage of soybean meal consumption by animal groups in the United States, Asian Subcontinent, Americas (Non-US), North Asia, Middle East and North Africa, Southeast Asia, and Sub-Saharan Africa from 2017-2018 (USB, 2019. United Soybean Board Supply & Disappearance. USB Market View Database (n.d.). Available at: https://marketviewdb.centrec.com/sd/. Accessed: 3/3/2021).

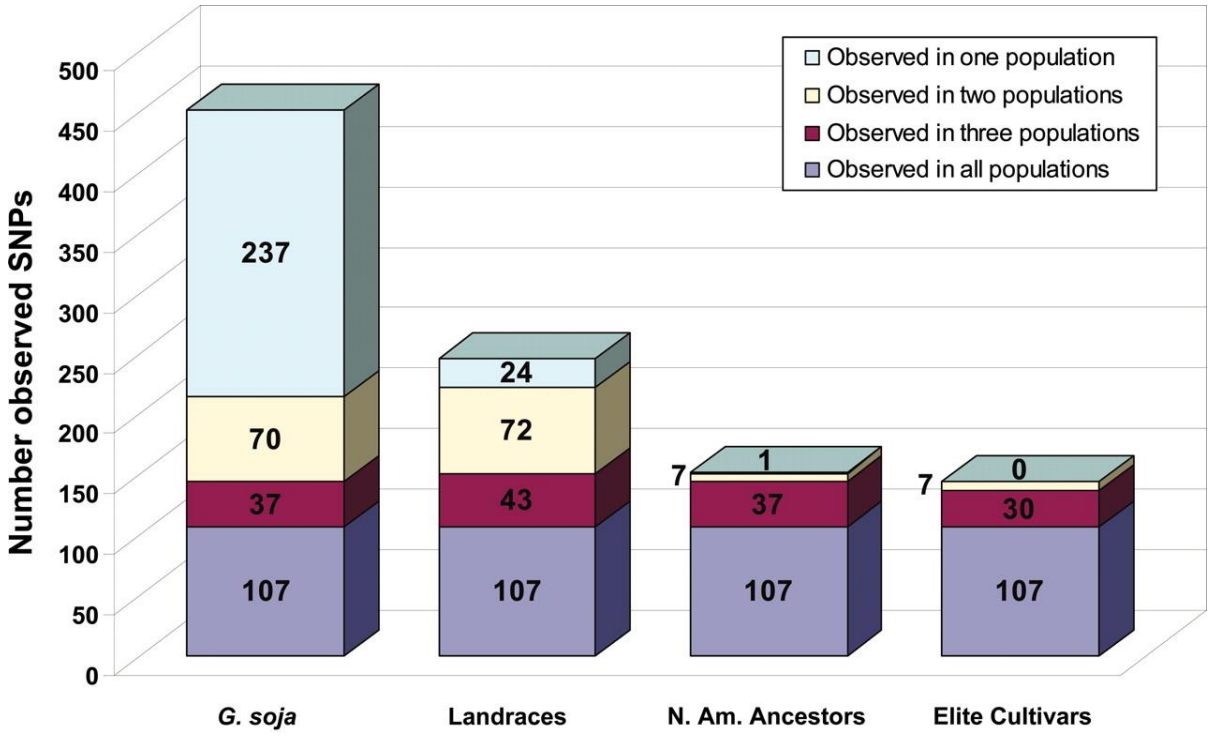

Figure 1-2. Number of observed SNPs across four studied groupings consisting of *G. soja* accessions, *G. max* Asian Landraces, North American ancestors, and elite cultivars (Hyten et al., 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences 103:16666–16671. doi: 10.1073/pnas.0604379103).

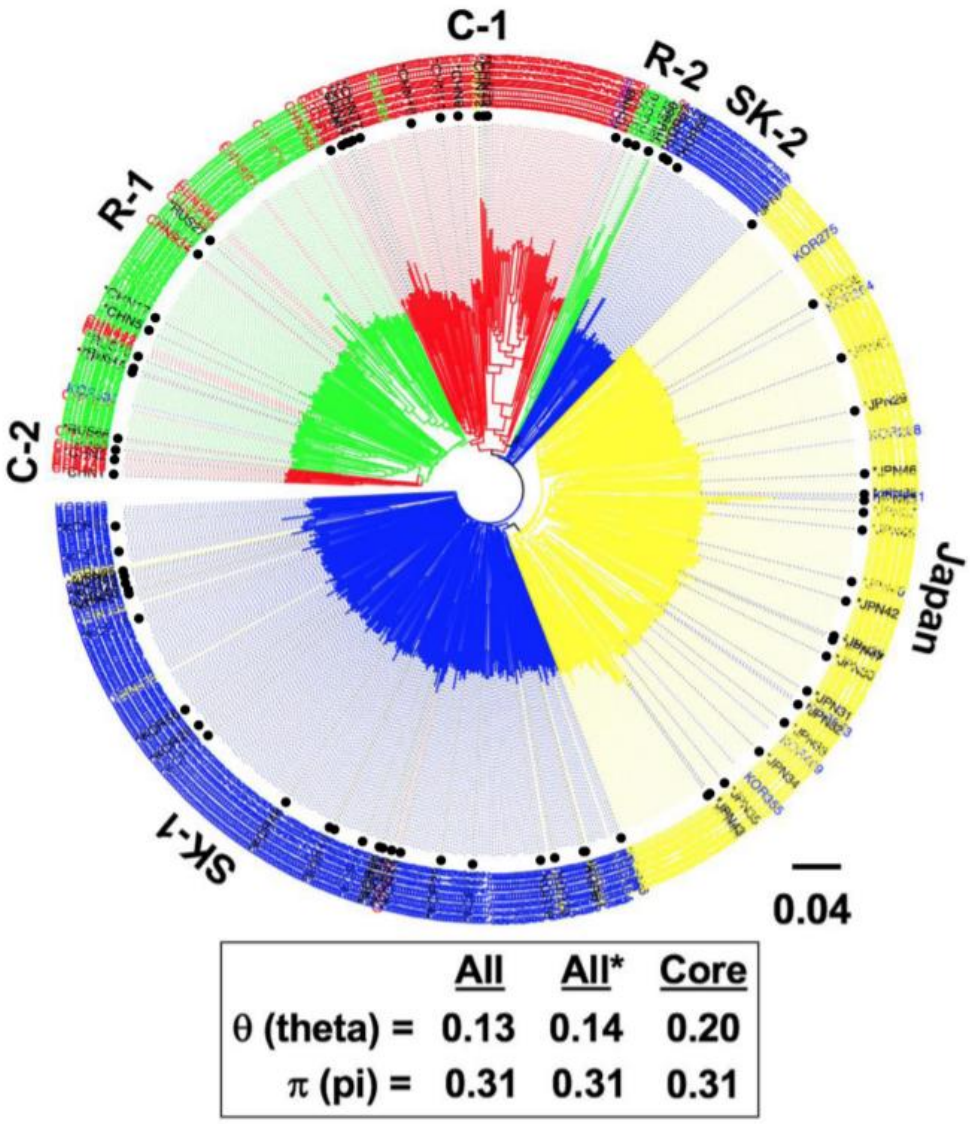

Figure 1-3. A phylogenetic tree representing the abbreviated USDA *G. soja* PI collection and the *G. soja* collection, along with the total (Θ) and average (π) nucleotide diversity estimates (La et al., 2019. Characterization of Select Wild Soybean Accessions in the USDA Germplasm Collection for Seed Composition and Agronomic Traits. Crop Science 59:233–251. doi: 10.2135/cropsci2017.08.0514).

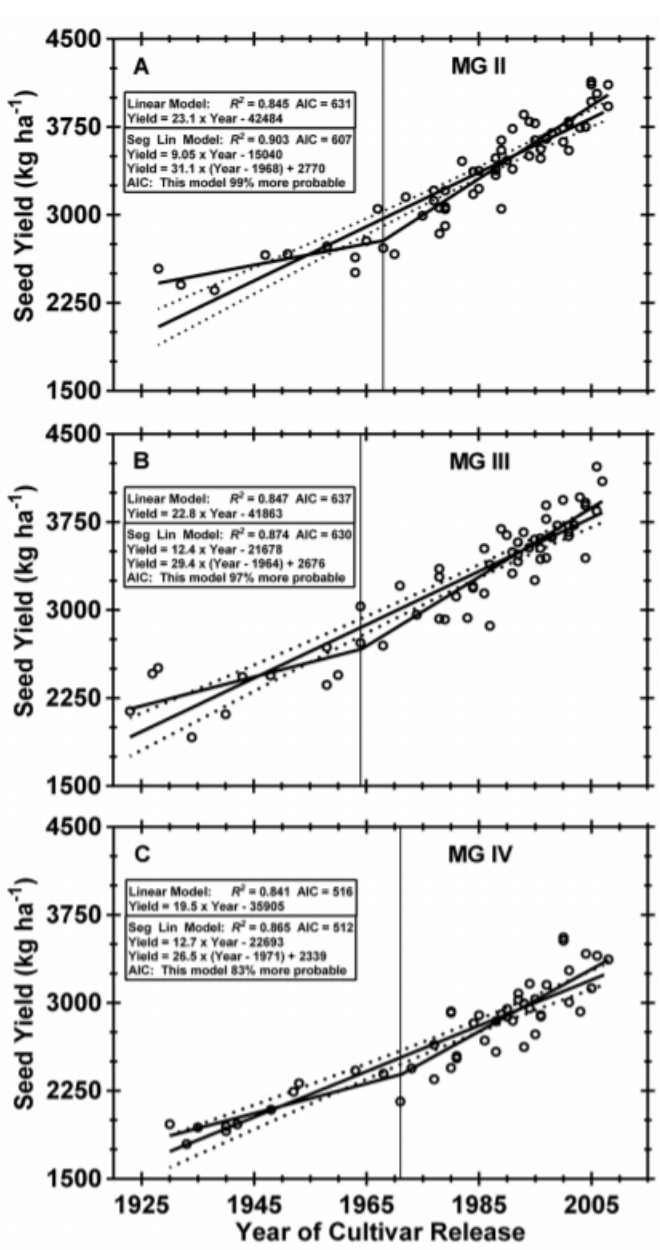

Figure 1-4. Increase in seed yield (kg ha^{-1}) from 1925 to 2005 in soybean maturity group II, III, and IV (Rincker et al., 2015. Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. Crop Science 54:1419–1432. doi: 10.2135/cropsci2013.10.0665).

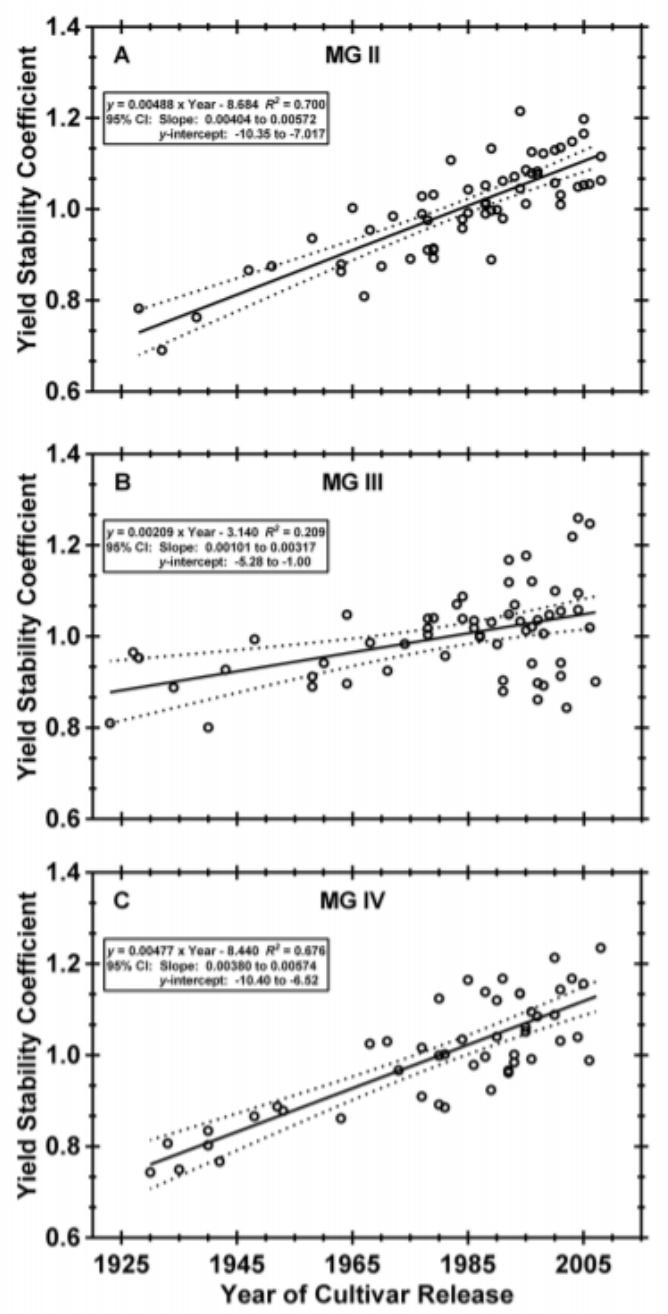

Figure 1-5. Seed oil and protein (g kg^{-1}) from 1925 to 2005 in soybean maturity group II, III, and IV (Rincker et al., 2015. Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. Crop Science 54:1419–1432. doi: 10.2135/cropsci2013.10.0665).

Table 1-1. Estimated rates of genetic gain of five agronomic traits and three end-use quality traits of soybean cultivars from maturity groups V, VII, and VII released from 1928 – 2008 (table modified from Boehm et al., 2019. Genetic Improvement of US Soybean in Maturity Groups V, VI, and VII. Crop Science 59:1838–1852. doi: 10.2135/cropsci2018.10.0627).

Trait	MG	*b*	± SE	R^2
Agronomic traits†				
Yield (kg ha^{-1} yr^{-1})	V	17.537***	1.307	0.841
	VI	13.503***	1.534	0.734
	VII	10.282***	1.139	0.765
Lodging (score yr^{-1}‡)	V	−0.011***	0.003	0.282
	VI	−0.009***	0.002	0.252
	VII	−0.013***	0.002	0.554
Plant height (cm yr^{-1})	V	−0.118NS§	0.102	0.037
	VI	−0.032NS	0.044	0.019
	VII	−0.061NS	0.061	0.037
Days to maturity (d yr^{-1}¶)	V	0.054NS	0.030	0.084
	VI	−0.002NS	0.033	0.001
	VII	−0.022NS	0.019	0.050
100-seed weight (g yr^{-1})	V	0.009NS	0.014	0.009
	VI	−0.019NS	0.016	0.049
	VII	−0.012NS	0.008	0.074
End-use quality traits				
Seed protein (g kg^{-1} yr^{-1}#)	V	0.022NS	0.095	0.002
	VI	−0.234**	0.070	0.270
	VII	−0.092NS	0.052	0.110
Seed oil (g kg^{-1} yr^{-1}#)	V	−0.026NS	0.072	0.004
	VI	0.074NS	0.059	0.053
	VII	0.041NS	0.057	0.021
Seed quality (score yr^{-1}††)	V	−0.005***	0.001	0.343
	VI	−0.001NS	0.001	0.003
	VII	−0.001NS	0.001	0.126

, * Significant at the 0.01 and 0.001, probability levels, respectively.

† Estimates were based on best linear unbiased predictor (BLUP) trait values generated from replicated yield trials conducted in 27 total southern US environments from 2010 to 2011. The BLUP trait value of each cultivar was regressed on the cultivar release date to estimate annualized genetic gains made over eight decades for each trait within each MG.

‡ Lodging score is assigned visually from 1 (all plants erect) to 5 (all plants prostrate to the ground).

§ NS, not significant.

¶ Days to maturity start on 1 September with positive BLUP values indicating later maturing cultivars and negative values indicating earlier maturing cultivars.

Protein and oil concentrations are expressed on a 130 g kg^{-1}moisture basis.

Chapter II

Fine Mapping and Candidate Gene Identification of a Soybean Seed Protein and Oil QTL from a Wild Soybean Accession

Abstract

Soybean [*Glycine max* (L.) Merr] cultivars have low genetic variation due to domestication, founder events, and selection strategies for modern plant breeding. Therefore, there is a need to introduce genetic diversity into soybean cultivars for long-term improvement of agronomic and seed compositional traits. In both public and private soybean breeding programs, the introgression of wild soybean (*Glycine soja* Siebold & Zucc.) genes has been utilized to incorporate novel genetic diversity. In our study, 3,015 single $F_{4:9}$ soybean plants were genotyped for nine genotype-by-sequencing markers from a previous genetic mapping study on recombinant inbred lines (La, 2018) to create two residual heterozygotes derived near isogenic lines (RHD-NIL) populations. The first RHD-NIL population was selected for a novel oil quantitative trait loci (QTL) on chromosome 8 and the second RHD-NIL population was selected for a novel protein QTL on chromosome 14. Both novel QTL derived from the wild soybean accession PI 593983. The objective of this research is to validate these QTL, reduce the QTL interval, and fine map the two novel QTL for candidate gene identification. Single marker analysis and linkage analysis was conducted using SoySNP6K BeadChip markers for QTL validation. The chromosome 8 oil QTL was not advanced for fine mapping because the QTL was not validated in a subsequent field and greenhouse study. Whole genome resequencing has been leveraged to reduce the QTL from 16.5 Mbp to approximately 4.6 Mbp and to fine map 50 high protein RHD-NIL, which have segregated for the validated chromosome 14 QTL to permit

candidate gene identification. A total of 55 potential candidates was identified in a physical interval of 8,059,955 to 12,648,760 bp. Our results may provide a better insight of utilizing wild soybean as a source of genetic diversity for soybean cultivar improvement.

Introduction

Soybean [*Glycine max* (L.) Merr] is one of the most valuable crops in the world due to the high protein and oil of its seed, which have uses as feed for livestock, a good source of protein and oil for human health, and as a biofuel stock (Masuda et al., 2009). The world total soybean production in 2019 was approximately 334 million metric tons (FAOSTAT, 2020). The increased use of soybean meal in animal feed as a protein source has been a major driving force in soybean production (Dei, 2011). Soybean meal is important for animals such as poultry, swine, aquaculture, and beef. Fifty-three percent of soybean meal sold in the United States was used in feed for poultry, 29% for swine feed, 8% for aquaculture, 7% for other animals, 2% for dairy, <1% for cattle feed, and <1% for companion animals (USB, 2019). Soybean oil is primarily used for human consumption as cooking oil, mayonnaise and salad dressing, and can also be used for industrial materials such as cement components, construction materials, electrical insulation, plastic, paint, mineral oils, and numerous applications (Hammond et al, 2005). Soybean oil is also a useful source for feed-grade fat (Dei, 2011).

Soybean cultivars have relatively low genetic variation due to evolutionary events such as domestication, founder events, and selection which can create genetic bottlenecks that can decrease genetic diversity, change allelic frequencies, increase linkage disequilibrium (LD), and eliminate rare alleles (Halliburton, 2004). Hyten et al. (2006) studied four populations and

reported a decrease of nucleotide diversity (π) from 2.17 x 10^{-3} in wild soybeans to 1.47 x 10^{-3} in landraces, to 1.14 x 10^{-3} in North American ancestors, and to 1.11 x 10^{-3} in elite cultivars. Li et al. (2013), Zhou et al. (2015), and Valliyodan et al. (2016) all reported similar nucleotide diversity levels, which indicate the bottleneck effects in soybean domestication. Hyten et al. (2006) also reported that 50% of the genetic diversity and 81% of the rare alleles have been lost during domestication and that 60% of the genes show significant changes in allelic frequency.

The two major soybean seed products are protein meal and oil, and soybean typically averages around 40% seed protein content and 20% seed oil content on a dry weight basis (Wilson, 2004). The inverse relationship between protein and oil is believed to be: 1) associated with environmental variability and genotype difference, 2) a single gene controlling multiple traits (pleitrophy), 3) tightly linked genes with different effects (Hymowitz, 1972; Chung et al., 2003; Leamy et al., 2017), and 5) energy cost partitioning between protein and oil (Egli and Bruening, 2007). Breeding for increased yield in soybeans has caused seed compositional traits such as seed protein content to decrease (Smith et al., 2004). Rincker et al. (2014) reported a positive increase in genetic grain yield and seed oil and a decrease in seed protein over 80 years (1925-2005) of breeding in soybean maturity groups II, III, and IV. Boehm et al. (2019) had a similar study but in US soybean maturity groups V, VI, and VII from 1930 - 2010 and reported a positive linear rate of genetic yield and seed oil improvement and a negative linear rate of genetic seed protein content. Brzostowki et al. (2017) suggested that it is important to evaluate a QTL across multiple genetic backgrounds and environment because of its complex relationships between seed composition and seed yield before incorporating it into a breeding program. Chung et al. (2003), Warrington et al. (2015), and Filho et al. (2015) all reported a negative correlation between seed yield and seed protein content. The negative correlation between seed protein and

seed oil, and seed protein and yield, make it challenging to develop a soybean cultivar that has increase seed protein and seed oil content as well as increased yield. Breeding for higher seed protein, seed oil, and yield in soybean germplasm can be difficult due to the negative correlation between seed protein and yield, seed protein and seed oil, and the positive correlation between seed oil and yield (Rincker et al., 2014, Wilson, 2004).

When crossing *G. soja* with *G. max*, undesirable traits from *G. soja* are often present in direct progeny, such as late flowering, hard seed coat, poor lodging, small seed size, pod shattering, and black color seeds (Carter et al., 2004, Liu et al., 2007). Many desirable genes from *G. soja* are thought to be linked to undesirable traits, making breeding with *G. soja* both time and resource intensive (Carter et al., 2004). Rare alleles are often lost during domestication or due to founder events. Such alleles have largely untapped potential for soybean improvement (Hyten et al., 2006). Previous studies from wild soybean or populations from crosses between *G. soja* and *G. max* can identify new potential genes, alleles, and quantitative trait loci (QTL) for genetic and agronomic improvements of traits such as yield and maturity (Li et al., 2008), soybean cyst nematode (Zhang et al., 2017), seed yield (Concibido et al., 2003), linolenic acid content (Pantalone et al., 1997), and seed protein content (Diers et al., 1992).

Currently there are 248 and 327 QTL associated with seed protein and seed oil content, respectively, as reported in the Soybean genetics and genomics database (https://www.soybase.org/.org, accessed on 11/12/2020). Many of these QTL were discovered through linkage analysis which requires F_2 generation, backcross, or RIL derived from original biparental crosses (Leamy et al., 2017). The first two major seed protein/oil QTL were discovered on chromosome (Chr.) 15 and 20 by Diers et al., (1992) from a cross between the *G. soja* accession PI 468916, a high protein wild soybean from Liaoning, China, and the *G. max* line

A8-3356022, a maturity group III experimental line from Iowa State University. The *G. soja* allele for the most significant marker from Chr. 20 and Chr. 15 had an increase in seed protein of 2.4% and 1.7%, respectively. These two QTL were later confirmed from the Soybean Genetics Committee (https://www.soybase.org/) and have been named cqSeed protein-001 (Fasoula et al., 2004) and cqSeed protein-003 (Nichols et al., 2006) for Chr. 15 and Chr. 20, respectively. Patil et al. (2018) studied an interspecific mapping population, consisting of 188 $F_{7:8}$ RIL, from a cross between the cultivar Williams 82 and a *G. soja* accession PI 483460B and identified five QTL for seed protein content on Chr. 6, 8, 13, 19, and 20 and nine QTL for seed oil content on Chr. 2, 7, 8, 9, 14, 15, 17, 19, and 20 by composite interval mapping using bin markers. Two significant seed protein loci were reported on Chr. 20 and one seed oil locus was identified on Chr. 5 using GWAS (Patil et al., 2018). Zhang et al. (2019) used a combination of linkage and GWAS analysis which identified four significant SNP loci regions distributed on Chr. 2, 6, 9, and 20 for seed protein and oil. The QTL on Chr. 20 explained the highest proportion of the phenotypic variance (7.27 to 9.39) and additive effect (0.56 to 0.75). All the QTL intervals reported either overlapped with or were close to, regions reported in previous studies (Diers et al., 1992; Qi et al., 2011; Tajuddin et al., 2005; Le et al., 2013; Pathan et al., 2013; Patil et al., 2018; Seo et al., 2019). Warrington et al. (2015) studied the Benning x Danbaekkong population and identified QTL for seed protein and amino acid on Chr. 14, 15, 17, and 20, and mapped Chr. 20 which explained 55% of the phenotypic variation and contains the *G. soja* Danbaekkong allele.

A novel seed protein QTL on Chr. 14 and seed oil QTL on Chr. 8 was detected in a previous study conducted by La, (2018) using a recombinant inbred line (RIL) population created from a single F_2 by crossing Osage x PI 593983 (a wild soybean line). Here we report on

two residual heterozygotes derived near isogenic lines (RHD-NIL) populations derived from two entries of the original RIL mapping population. The overall objective of this study were to 1) validate a seed protein QTL on Chr. 14, 2) validate a seed oil QTL on Chr. 8, 3) validate the RHD-NIL as true near isogenic lines (NIL), 4) reduce the initial QTL, and 5) fine map both QTL to allow candidate gene identification.

Materials and Methods

Population Development and Field Experiments

The parental lines, Osage [*Glycine max* (L.) Merr.] and PI 593983 (*G. soja* and Zucc.) were crossed in North Carolina in 2011. The F_1 generation was grown at a USDA-ARS winter nursery in Isabela, Puerto Rico during the winter of 2011/2012. In the summer of 2012, the F_2 generation was grown in Columbia, MO. A single F_2 plant was selected and F_3 seeds were harvested from this single plant. During the summer of 2013, 338 F_3 plants were planted in single row and were harvested individually. 338 F_4 single row plots with two replications were grown at Bradford Research Center, Columbia, MO in the summer of 2014. One F_4 plant was harvested within each RIL for all plots for a total of 338 F_4 derived plants. During the winter of 2014/2015, $F_{4:5}$ RIL were planted in Florida, USA for seed increase. In the summer of 2015, 181 out of 338 $F_{4:6}$ was randomly selected and planted at Bay Farm Research Facility, Columbia, MO. 174 $F_{4:7}$ RIL were planted during the summer of 2016 at three locations: Bay Farm Research Facility, Columbia, MO; Lee Greenley Memorial Jr. Research Center, Novelty, MO; and Hundley-Whaley Research Center, Albany, MO. During the summer of 2017, 164 $F_{4:8}$ RIL were planted at Bradford Research Facility, Columbia, MO and Lee Greenley Memorial Jr.

Research Facility, Novelty, MO. 13 RIL were selected for being heterozygous for multiple QTL regions and were grown at Bay Farm Research Facility, Columbia, MO in 2018. Due to a limited number of seeds, 121 $F_{9:10}$ NIL, derived from two $F_{4:9}$ RIL, were grown with two replications as hill plots (1-8 seeds per hill plot) in the summer of 2019 at Bay Farm Research Facility, Columbia, MO (19CLM) and Lee Greenley Memorial Jr. Research Facility, Novelty, MO (19NOV). In the summer of 2020, 53 $F_{9:11}$ NIL were grown as hill plots (25 seeds per plot) with two replications at Bay Farm Research Facility, Columbia, MO (20CLM) and Lee Greenley Memorial Jr. Research Facility, Novelty, MO (20NOV).

Greenhouse Experiment

During the winter of 2018/2019, 139 $F_{9:10}$ RHD-NIL derived from two $F_{4:9}$ RIL were planted at Ashland Greenhouse, University of Missouri, Columbia, MO, USA (18/19GH) with two replications. Two plants were planted and grown per pot for each RHD-NIL for a total of 278 pots. RHD-NIL were harvested, and seeds were bulked by individual pot for seed protein and seed oil analysis.

Genotyping Analysis

In 2018, approximately 3,015 single $F_{4:9}$ soybean plants were genotyped for 28 markers with five markers on the Chr. 14 protein QTL and four markers on the Chr. 8 oil QTL using multiplexed Next-Gen PlexSeq™ from AgriPlex Genomics (AgriPlex Genomics, Cleveland, OH, USA). Leaf tissues were collected from every plant in a 2ml tube and then lyophilized for 48 hours before samples were sent to AgriPlex. PlexSeq™ technology provides focused next generation sequencing analysis for any SNP of interest and the multiplex capabilities of PlexSeq™ provides data at a lower cost and in less time than single-plex approaches. AgriPlex

Genomic performed all library construction, Illumina sequencing, and genotype calling (AgriPlex Genomics).

AgriPlex used a in house software called PlexCall™ to call genotypes with AA as parent 1, BB as parent 2, HH as heterozygous, and RECOMB as recombination of the parent's alleles. Due to genotyping error by AgriPlex, 61% of the genotyping data were usable. After removing all missing data and errors, NIL were selected to cover all genotypic classes for the Chr. 14 protein QTL: 61 NIL were recombinant, 10 NIL was called AA, which is the genotype from the parent Osage, 10 NIL was called BB, which is the genotype from the parent PI 593983, and 2 NIL was called HH for the QTL on Chr. 14. For the QTL of Chr. 8, there were 66 total genotypes with 46 NIL marked as recombinant, 10 NIL called as AA, 8 NIL called as BB, and 2 NIL called as HH.

During the summer of 2019, 119 $F_{9:10}$ RHD-NIL were grown with two replications at 19CLM and 19NOV as hill plots. The 2019 hill plots varied from one to eight plants due to the limited number of seeds available. Young trifoliate were collected from every plant in the hill plots from 19CLM and bulked per plot during the V5 growth stage. DNA extraction was conducted from a modified Cetyl Trimethyl Ammonium Bromide (CTAB) method (Doyle and Doyle, 1987). Tissues were grinded with beads using a Mini-Beadbeater-96 (BioSpec Products, Bartlesville, OK, USA) for 30 seconds at 2,100 revolutions per minute (rpm). 1,000 mircoliter (uL) of CTAB extraction buffer was added to each sample, then vortex to re-suspend the powder, and finally incubated in water batch at 60 degree Celsius for 30 minutes. 500 uL of 24:1 Chloroform/isoamyl alcohol were added to each sample and samples were centrifuged for five minutes at 5,000 centrifugal force (G). 700 uL of supernatant from each samples were transferred to new tubes and 2.5 uL of 10 milligram/milliliter of RNase were added to each sample. 500 uL

of 24:1 Chloroform/isoamyl alcohol were added to each sample, samples were centrifuged for five minutes at 5,000 centrifugal force (G), and 500 uL of supernatant from each samples were transferred to new tubes. 800 uL of cooled isopropanol were added to each samples, samples were centrifuged for five minutes at 5,000 G, isopropanol were removed from each samples by pouring it out, 200 uL of 70% ethanol were added to each samples, samples were spun down in the centrifuge at 5,000 G, and finally ethanol were removed from each sample. Pellets were air dried over night and on the next day, 1,000 TE buffer were added to each sample. The modified CTAB method was used to extract high quality DNA suitable for BARCSoySNP6K BeadChip genotyping array (SoySNP6K) and whole genome resequencing. DNA samples were then submitted to the Soybean Genomics and Improvement Laboratory, USDA-ARS Beltsville, MD, for BARCSoySNP6K BeadChip Illumina Infinium genotyping array (Song et al., 2020).

Whole Genome Resequencing

A total of 53 RHD-NIL DNA samples were submitted to a commercial vendor, GENEWIZ for short-read whole genome sequencing at approximately 15x coverage. Genomic variations were identified with PGen, a genomic variation analysis workflow (Liu et al., 2016), which is large-scale next generation resequencing (NGS) data analysis of genomic variations workflow. PGen was used to efficiently facilitate large-scale NGS data analysis of genomic variations which is available in both a Linux version and a web-based implementation integrated within SoyKB (Joshi et al., 2014) and KBCommons (Zeng et al., 2019). *G. max* Williams 82 was the genotype, and the Wm82.a2.v1 assembly (Schmutz et al., 2010) available via Phytozome (Goodstein et al., 2010) was used as the reference genome for mapping. The workflow starts by accepting paired-end or single-end fastq reads as input and performs data quality checks as the first step using FastQC (Andrews, 2010). Only the filtered high- quality reads are later aligned

against the reference genome using BWA (Li et al., 2009). Picard Tools (Picard, 2018) was also used at this step to locate duplicate molecules and assign all reads into groups with the default parameters. After alignment, SNPs and indels were called using the Haplotype caller algorithm from the Genome Analysis Toolkit (GATK) (McKenna et al., 2010). Filtering criteria were defined in INFO field in vcf file, where QD stands for quality by depth, FS is Fisher strand values and MQ is mapping quality of variants. Detected variants were then filtered using the criteria "QD < 26.0 || FS > 60.0 || MQ < 40.0" for SNPs and "QD < 26.0 || FS > 200.0 || MQ < 40.0" for indels.

SoySNP6K Data and Whole Genome Resequencing Data Quality Control

Alleles were called using the software GenomeStudio v2.0.5 (Illumina, San Diego, CA, USA). Quality control was conducted in TASSEL version 5.0 (Bradbury et al., 2007) with adjusted parameter following Heim et al. (2017) by removing markers greater than 80% heterozygous and removing RHD-NIL that have greater than 10% missing data. ABH parental calls were conducted in TASSEL version 5.0, where A represents parent one allele homozygote (Osage), B represents parent two allele homozygote (PI 593983), and H represents heterozygous (AB heterozygote). Genotypic data was extracted from TASSEL version 5.0 and imported into RStudio version 1.2.1335. The package 'ABHgenotypeR' (Furuta et al., 2017) was used to impute missing genotypes based on flanking alleles for error correction with the adjusted parameter of maxHapLength = 3 based on the work from Zhu et al. (2021), resulting in 2,966 markers.

A total of 431,738 SNP were called on Chr. 14 from the whole genome resequencing (WGR) data. An adjusted strict quality control following Heim et al. (2017) were imposed in Tassel version 5.0 to call parental genotypes. The minimum SNP count was set at 30 and SNP

greater than 80% heterozygous and less than 10% allelic frequency were removed. SNP were filtered again with the minimum SNP count at 35 out of 55 sequence, maximum allelic frequency of 90% and minimum allelic frequency at 10%. The function 'homozygous genotype' was used to remove all heterozygous allele calls. The function 'thin site by position' was used to remove a SNP at every 2000 base pair. LD KNNi imputations were conducted and ABH parental calls were conducted in Tassel version 5.0. Genotypic data were imported into RStudio version 1.2.1335 and the package 'ABHgenotypeR' (Furuta, et al., 2017) was used for error correction using the adjusted parameter of maxHapLength = 5 based on the work from Zhu et al. (2021), resulting in 11,836 SNP markers.

Seed Protein and Oil Analysis

In this study, 278 RHD-NIL from 18/19GH and 415 samples from the 2019 field study from two locations, 19CLM and 19NOV were collected for protein and oil analysis of the two QTL on Chr. 8 and Chr. 14. A total of 100 RHD-NIL of the protein Chr. 14 QTL from two locations, 20CLM and 20NOV, were collected for protein and oil analysis in 2020. Approximately 20 ml of whole seeds were allocated from each RHD-NIL and grounded using a Perten laboratory Mill 3600 grinder (Perten Instruments, Hägersten, Sweden). Samples were analyzed for protein and oil content on a dry weight basis via near-infrared spectroscopy (NIRS) using a Perten model DA 7250 (Perten Instruments, Hägersten, Sweden). NIRS calibrations were originally developed and are updated every year by Perten Instruments and the University of Minnesota technical staff. The 2018/2019 greenhouse study were analyzed using the 2018 ground NIRS calibration, the 2019 field study were analyzed using the 2019 ground NIRS calibration, and the 2020 field study used the 2020 ground NIRS calibration (Supplementary Table 2-1).

Statistical Analysis of Phenotypic Data

Statistical analysis was conducted in RStudio version 1.2.1335 (RStudio Team) using the function 'aov' to compute analysis of variance (ANOVA). Single marker analysis using the SNP called from the BARCSoySNP6K BeadChip genotyping array was used for validating the Chr. 8 oil QTL and Chr. 14 protein QTL. Genetic similarity was calculated in TASSEL version 5.0 using the 'distance matrix' function to validate the Chr. 14 RHD-NIL as true NIL. ANOVA and broad-sense heritability on an entry mean basis were calculated using phenotypic values of the two replicated lines in each environment. The ANOVA statistical model is shown below:

$$y_{ijk} = \mu + G_i + G_iE_j + E_j + R_{kj} + e_{ijk}$$

where y_{ijk} represents phenotype of in the *i*th genotype under the *k*th environment being the *k*th replication within the *j*th environment, μ represents the population mean, G_i represents the *i*th genotype, G_iE_j represents the *i*th genotype by *j*th environment interaction, E_j represents the environmental effect, R_k is the *k*th replication within the *j*th environment, and e_{ijk} represents the residual effects (Fehr, 1991; Bernardo, 2020).

Broad-sense heritability on an entry-mean basis was estimated using the formula below:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GE}^2}{E} + \frac{\sigma_e^2}{RE}}$$

where h^2 indicated broad-sense heritability on an entry-mean basis, σ_G^2 is the genotypic variance, σ_{GE}^2 is the genotype x environment variance, *E* is the number of environments. σ_e^2 is the error variance, and *R* is the number of replications (Falconer and Mackay, 2009; Fehr, 1991; Bernardo, 2020).

Significant recombination regions were determined by using a modified Best Linear Unbiased Prediction mixed-linear-model (Bernado, 1994; Panter and Allen, 1995) in RStudio version 1.2.1335 (RStudio Team) using the function 'lmer'. The mixed-linear-model is described below:

$$y_{ik} = \mu + M1 + M2 + M3 + M4 + M5 + M6 + M7 + M8 + E_j + R_{kj} + e_{ik}$$

where μ is the mean, $M1$ is marker that represents the first recombination region, $M2$ is the second recombination region, $M3$ is the third recombination region, $M4$ is the fourth recombination region, $M5$ is the fifth recombination region, $M6$ is sixth recombination region, $M7$ is the seventh recombination region, $M8$ is the eight recombination region, E_j is the environmental effect, R_{kj} is the *k*th replication within the *j*th environment effect, and e_{ik} represents the residual effect. $M1$ – $M8$ are fixed effects and E_j and R_{kj} are random effects (Bernado, 1994; Panter and Allen, 1995).

Genetic Map and Linkage Analysis

The genetic map and QTL mapping for protein was created in RStudio version 1.2.1335 (RStudio Team) using the package 'qtl' (Broman et al., 2003; Broman, 2011). There were 2,962 SNP6k markers across 20 chromosomes after dropping markers that were not present on more than 50 RHD-NIL. A total of 93 SNP6k markers were present on Chr. 14 and used for QTL mapping. The function 'scanone' and using the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) and Haley-Knott regression method (regression of the phenotypes on the multipoint QTL genotype probabilities), as described by Haley and Knot (1992), was used for interval mapping on the Chr. 14 protein QTL. Due to the low density of markers on Chr. 14,

interval mapping was unable to narrow the QTL to a manageable region for candidate gene prediction.

A genetic map of the WGR SNP was created in RStudio version 1.2.1335 (RStudio Team) using the package 'qtl' (Broman et al., 2003; Broman, 2011) for QTL mapping. 11,836 SNP were reduced to eight SNP using the functions 'findDupMarkers' and 'drop.markers'. The 'findDupMarkers' function identify sets markers that are in linkage or are genetically identical. The eight SNP represents eight different physical base pair regions of recombination events on Chr. 14. The 'drop.markers' keeps the first marker of the recombination regions and drops the remaining markers that are in linkage. QTL mapping was conducted using the function 'cim' for composite interval mapping on the Chr. 14 protein QTL with the number of marker covariates set at 5, a mapping interval of 10 centimorgan (cM), EM as the mapping method, and error probability of 0.001.

Candidate Genes Selection

Gene models and gene annotations were extracted from Soybase (www.soybase.org, accessed on 3/01/2021). Potential candidate genes were selected based on gene ontology (GO) biology descriptions, which was obtained from TAIR v 10 (03/27/14), and EuKaryotic Orthologous Groups (KOG) descriptions from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax).

Results

Phenotypic Analysis of Seed Protein and Oil

The phenotypic analysis for oil and protein content were conducted for all environments (19CLM, 19NOV, 20CLM, and 20NOV), including combining the two field seasons together (CLM&NOV) and combining all environments which includes the 2018/2019 greenhouse study (Combined). The 2018/2019 greenhouse study was left out of the CLM&NOV because the mean had too large of a margin under a mean-separation test to be grouped together with the field studies and the descriptive phenotypic values of seed oil and protein in this study are presented in Table 2-1. The oil content across environments range from 18.0 to 22.1%, while the combined field seasons ranged from 17.4 to 18.9% and the greenhouse study ranged from 20.1 to 22.1%. The coefficient of variation (CV) for oil content ranged from 1.82 to 3.10% across all environments.

The skewness and kurtosis of distribution are available in Table 2-1. In the environments 19CLM and 19NOV for protein content, the kurtosis absolute value is above one. While the rest of the environments, CLM&NOV, and combined environments, the kurtosis absolute value is near one. The absolute value for skewness was below one for both oil and protein content. The absolute kurtosis value for oil content was above 0.50 for CLM&NOV and 18/19GH were above and the rest of the environments were below 0.50. These results indicated that the oil and protein content follow a continuous and normal distribution. The RHD-NIL population were selected based on alleles that are segregated for either the *G. max* parent or the *G. soja* parent in the QTL, thus suggesting a random population structure to move forward with fine mapping.

A Pearson correlation analysis was conducted for phenotypic values of the RHD-NIL in each environment (Table 2-2). A highly significant correlation existed between seed oil and seed protein content between environments ($P < 0.001$). These results suggested that increasing seed protein content will decrease seed oil content. Therefore, breeding for increased seed protein and seed oil content is extremely difficult.

An ANOVA and broad-sense heritability on an entry-mean basis were conducted for the following environments: 19CLM, 19NOV, 20CLM, and 20NOV (Table 2-3). The genotypic variance explained for seed protein content was the highest variance explained at 2.27 and was significant at $P<0.001$, with environmental variance at 1.86. For oil content, the environmental variance was the highest at 480.91, followed by the genotype variance at 4.21. Genotype and environment were both significant at $P<0.001$ for seed oil content. The entry-mean based heritability (h^2) for seed protein content was 0.72 and seed oil content was 0.69. The results from the ANOVA suggested that the genotypes from the RHD-NIL had a bigger impact on the level of seed protein content and the environment had a bigger impact on the level of seed oil content.

Validation of the Chr. 14 Protein QTL and the High Protein RHD-NIL Population

Two QTL were detected from the initial RIL population (La, 2018) with a seed oil QTL on Chr. 8 and a seed protein QTL on Chr. 14. SoySNP6K markers were used to validate the two QTL using the 2019 field dataset. The Chr. 8 oil QTL was not validated based on single marker analysis using 62 SoySNP6K markers (Figure 2-1). The Significant threshold for the Chr. 8 oil QTL was at 3.09 $-\log_{10}(P)$ based on a Bonferroni Correction. The highest $-\log_{10}(p)$ value of protein was approximately 0.33 and oil was approximately 0.69 which is extremely below the Significant threshold. The Significant threshold based on a Bonferroni Correction for the Chr. 14 protein QTL was at 3.27 $-\log_{10}(P)$. 71 of 93 SoySNP6K markers were significant with $-\log_{10}(p)$

values ranging from 3.32 to 6.40 for seed protein content and 64 SoySNP6K markers were significant with $-\log_{10}(p)$ values ranging from 3.33 to 4.72 for seed oil content (Figure 2-2). These findings suggested that Chr. 8 oil QTL was detected as a false positive from the RIL mapping population study (La, 2018) and was not continued for further analysis. While Chr. 14 protein QTL was validated and was moved forward for fine mapping.

The next step was to validate that the high protein RHD-NIL are in fact true NIL. A genetic similarity test indicated that the high protein RHD-NIL are genetically similar (Figure 2-3). Between the Osage and PI 593983, they were 49% genetically similar and individual RHD-NIL compared to the parental lines are genetically similar ranging from 71 to 73% for Osage and 68 to 71% for PI 593983. Between individual RHD-NIL, they ranged from 96 to 99% genetically similar. These results indicated that the parental lines are genetically different from one another. Individual RHD-NIL inherited alleles from both parental lines across 20 chromosomes. Between individual RHD-NIL, they are extremely genetically similar which validated the high protein RHD-NIL as true NIL. Thus, suggesting that the seed protein QTL on Chr. 14 is one of a few genomic locations across 20 chromosomes that are still segregating alleles between parents. The high protein RHD-NIL can now be treated at true NIL for fine mapping to identify candidate genes.

Fine-mapping the Chr. 14 Protein QTL

In this study, the initial Chr. 14 protein QTL was approximately 16.5 million base pairs (Mbp) from the RIL population (Figure 2-4A). SoySNP6K markers was used to fine map the QTL but due to the limited number of recombination events, there were gaps present between markers Gm14_4728306 and Gm14_20110020 spanning a physical distance of approximately 15.4 Mbp.(Figure 2-4B). The Chr. 14 protein QTL (*qPro*-14-1) in the RHD-NIL population

ranged from 5,509,372 to 14,976,378 base pairs (bp) for a total range of approximately 9.5 Mbp using WGR data. A total 11,836 SNP from WGR data broke into eight recombination regions representing eight recombination events on Chr. 14 (Figure 2-4C). The eight recombination regions corresponding to the Chr. 14 QTL ranged from Gm14_5509372 to Gm14_6485179 (*rr*-14-1), Gm14_6487608 to Gm14_7138691 (*rr*-14-2), Gm14_7141628 to Gm14_7453099 (*rr*-14-3), Gm14_7455192 to Gm14_8048870 (*rr*-14-4), Gm14_8059955 to Gm14_9506311 (*rr*-14-5), Gm14_9508613 to Gm14_12648760 (*rr*-14-6), Gm14_12655776 to Gm14_14976378 (*rr*-14-7), and Gm14_14976378 to Gm14_44140803 (*rr*-14-8) (Table 2-4). The eight recombination regions reduced the QTL interval to approximately 4.6 Mbp. Gm14 denotes Chr. 14, and the following numbers is the physical location of the recombination region in base pair. The genetic position of the eight seed protein and seed oil recombination regions from one to eight are 0.00, 3.32, 5.46, 16.06, 23.41, 25.57, 36.20, and 37.26 centimorgan (cM), respectively (Table 2-4).

The phenotypic variance (R^2) explained for seed oil content ranged from 2.82 to 4.81% (Table 2-4). RHD-NIL with the *G. soja* allele (TT) at *rr*-14-5 saw a decrease of seed oil content of 0.40% (Figure 2-5A) and the *G. soja* allele (GG) at *rr*-14-5 also a decrease of oil content of 0.44% (Figure 2-5B) from the environment CLM&NOV. The difference for oil content between the *G. max* and *G. soja* allele in the greenhouse study at *rr*-14-5 and *rr*-14-6 was 0.37% (Figure 2-5C) and 0.42%, respectively (Figure 2-5D).

Two recombination regions for seed protein content were significant based on the F-value (Table 2-4). *rr*-14-5 and *rr*-14-6 had F-values of 5.60 ($P < 0.05$) and 7.03 ($P < 0.01$), respectively. *rr*-14-3 was also significant ($P < 0.1$) with an F-value of 3.73. The phenotypic variance (R^2) explained for protein content ranged from 10.47 to 17.99% with *rr*-14-3 at 16.43%, *rr*-14-5 at 12.61%, and *rr*-14-6 at 16.16%. The *G. soja* allele (TT) at *rr*-14-5 increased protein

content at an average of 0.65% from the *G. max* allele (CC) (Figure 2-6A), while the *G. soja* allele (GG) increased protein content at an average of 0.72% compared to the *G. max* allele (TT) at *rr*-14-6 (Figure 2-6B) from the environment CLM&NOV. In the greenhouse study, *rr*-14-5 increased by 1.57% (Figure 2-6C) and *rr*-14-6 increased by 1.75% for protein content (Figure 2-6D). This analysis further fine mapped the QTL region to *rr*-14-5 and *rr*-14-6 for protein content and it spans from 8,059,955 to 12,655,776 bp. Candidate genes can be predicted from this approximately 4.6 Mbp region for the increase in protein content.

Candidate Gene Prediction

A total of 223 genes (Glyma.Wm82.a2.v1) are present in *rr*-14-5 and *rr*-14-6 were retrieved from Soybase (http://www.soybase.org). In *rr*-14-5 (8059955 – 9506311 bp), 24 out of 100 genes with protein transport, amino acid transport, amino acid biosynthesis process, seed development, and protein catabolic process were selected as candidate genes based on GO annotations (Table 2-5). For *rr*-14-6 (9508613 –12648760 bp), 26 out of 123 genes were selected as candidate genes (Table 2-6). A total of 50 genes were identified as candidate genes in *rr*-14-5 and *rr*-14-6, and are within the physical interval of 8,059,955 to 12,648,760 bp.

Discussion

In this study, we leveraged SNP data from WGR data for our statistical analysis to fine mapping our QTL. Existing BeadChip arrays, such as the BARCSoySNP6K BeadChip Illumina Infinium genotyping array (Song et al., 2020) which is a subset derived from the BARCSoySNP50K BeadChip Illumina Infinium genotyping array (Song et al., 2013), is a strong tool for genetic research that has been used to identify QTL and genes associated with

phenotypic traits like growth period (Liu et al., 2016), seed oil and fatty acids content (Priolli et al., 2019), seed protein content (Nascimento et al., 2018) and seed yield (Ye et al., 2018). In inbred lines, there are a limited number of recombination events which suggests that it is unnecessary to genotype lines with many markers for a biparental population (Song et al., 2020). In our study, due to genotyping error in 2018 by AgriPlex, our total sample size greatly decreased, which then affected the number of recombination events in our RHD-NIL population. This caused the SoySNP6K markers to not be able to finely map the Chr. 14 protein QTL due to limited genetic diversity and insufficient polymorphic markers. The advancement and lower cost of next-generation sequencing has become a strong tool in the field of genomics by allowing researchers to sequence whole genomes (Koboldt et al., 2013). Individual-based WGR obtains high-quality individual genotypes, which requires a high read depth to accurately identify SNP, short INDEL, and genotype calling (Nagasaki et al., 2015). NGS technology can generate thousands to millions of DNA sequences which can be leveraged to define genomic regions and increase SNP density and even identify molecular genetic causes for traits of interest (Park and Kim, 2016; Schaid et al., 2018). Patil et al. (2016) identified a major QTL on Chr. 3 that contains the salinity tolerance gene, *GmCHX1*, using WGR and SoySNP50K data. WGR SNP can be translated into functional markers and allows for further research on haplotype and SNP variation using WGR data (Patil et al., 2016). As NGS continues to advance and the cost continues to lower, researchers will be able to utilize this genomic tool for linkage analysis, fine mapping, gene cloning, and other scientific projects.

In our study, the seed protein QTL on Chr. 14 was validated by detecting an association between SoySNP6K markers with seed protein and oil content. The seed oil QTL on Chr. 8 was reported as a false positive QTL by single marker analysis using SoySNP6K markers. Multiple

environments, increased replications, larger sample size, and accounting for all QTL effects lowered the residual variance in the RIL mapping population and thus, detected a seed oil QTL on Chr. 8 (Falke et al., 2010). We suggest due a small sample size in the RHD-NIL population, high standard error probability, and high residual variance in the Chr. 8 RHD-NIL population would result in a low power of detection.

Near-isogenic lines are the preferred population for fine mapping because the genetic background is identical or nearly-identical between individual NIL, except for the targeted genomic region, which allows us to accurately model the effect of the QTL and by examining multiple NIL, it is possible to break up a large QTL interval into shorter intervals (Fridman et al. 2000; Jander et al. 2002; Uga et al.2013; Song et al. 2015). In our study, we were able to decrease the size of the initial QTL detected in the RIL population using a mixed-linear-model. Although we identified a very large number of polymorphisms (11,836), the limited recombination condensed to a single representative marker per recombination region which were used for analysis. We were able to reduce the Chr. 14 protein QTL to two recombination regions (*rr*-14-5 and *rr*-14-6) that are very significantly associated with the increase in seed protein content. Similar fine mapping approaches have been conducted using either single marker regression or haplotype analysis; Zhang et al. (2012) fine mapped a major flowering time QTL, *qFT6,* by performing haplotype analysis between every two markers and performed regression analysis of the haplotypes to the phenotypic data.

The initial novel Chr. 14 protein QTL in this study was detected in a RIL population from a previous study (La, 2018). Our study was to narrow the QTL interval for predictive gene identification and for breeding purposes. Multiple seed protein and oil QTL have been detected and studied on Chr. 5 (Pathan et al., 2013), Chr. 15 (Diers et al., 1992; Fasoula et al., 2004;

Pathan et al., 2013; Warrington et al., 2015), and Chr. 20 (Diers et al., 1992; Nichols et al., 2006; Patil et al., 2018). Warrington et al. (2015) identified a protein QTL on Chr. 14 with a phenotypic variance of 5% derived from Benning x Danbaekkong. Zhang et al. (2004) identified a QTL on Chr. 14 from a Kefong No.1 x Nanong 1138-2 and had a phenotypic variance of 12.4%. Many of the detect protein QTL on Chr. 14 have alleles derived from Asian landraces (Zhang et al., 2004; Warrington et al., 2015; Huang et al., 2020). In our study, the allele responsible for the increase in protein content was from a *G. soja* accession, PI 593983. The genetic diversity in *G. soja* is more diverse when compared to Asian landraces (Hyten et al., 2006). The phenotypic variation explained for protein content in our study was 12.61% for *rr*-14-5 and 16.16% for *rr*-14-6 and heritability h^2 was 0.72 for protein content and 0.69 for oil content. The region *rr*-14-5 is located from 8,059,955 to 9,506,311 bp and *rr*-14-6 physical interval is 9,508,613 to 12,648,760 bp.

In this study, a total of 50 protein candidate genes were identified which are located in the physical interval of 8,059,955 to 12,648,760 bp. At total of 24 candidates genes were identified within *rr*-14-5 and another 26 candidate genes within *rr*-14-6. These candidate genes were selected based on their gene ontology annotations from Soybase (www.soybase.org, accessed on 3/01/2021) related to protein transport, amino acid transport, seed development, amino acid biosynthesis, and protein catabolic process. In *rr*-14-5, three genes had biological functions of amino acid transport, *Glyma.*14G090200, *Glyma.*14G096200, and *Glyma.*14G096600, and *Glyma.*14G098100 was described as having a cellular modified amino acid biosynthesis. In *rr*-14-6, *Glyma.*14G102700 has a biological function described as aromatic amino acid family biosynthesis, *Glyma.*14G104800 regulated amino acid import, *Glyma.*14G105200 regulated amino acid export, and *Glyma.*14G105900 has a biological process

of amino acid transportation. These candidate eight genes could be responsible for the increase in seed protein content. These reported 50 genes can be considered as potential candidate genes for seed protein, but additional research is required to further narrow our candidate gene list to identify a causative polymorphism(s) within a specific gene(s).

Conclusion

A novel Chr. 14 QTL identified in our study was limited to a total of eight recombination events. The eight recombination events were broken into eight recombination regions which were used for statistical analysis to identify significant regions responsible for the increase in seed protein content. Potential candidate genes were selected based on protein and oil storage, protein transport, amino acid transport, seed development, amino acid biosynthesis, and protein catabolic process, for a total 50 genes. This novel QTL has the potential be used for the introgression of increased protein into cultivar traits. Further analysis needs to be conducted to reduced and validated predictive genes for potential candidate genes. The integration of using wild soybean germplasm as a source of genetic diversity is still fairly new because of the difficulties of working with wild soybeans. Our results indicates that utilizing wild soybean has the potential to be effective in introducing favorable alleles into elite soybean cultivar germplasms for seed compositional improvements.

References

Abdelghany AM, Zhang S, Azam M, et al (2019) Natural Variation in Fatty Acid Composition of Diverse World Soybean Germplasms Grown in China. Agronomy 10:24. doi: 10.3390/agronomy10010024

Andrews, S.: FastQC: a quality control tool for high throughput sequence data: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010

Agriplex Genomics – Innovating Focused Population Scale Genotyping Sequencing. (2019). Retrieved December 27, 2019, from Agriplexgenomics.com website: http://agriplexgenomics.com/

Akond M, Liu S, Boney M, et al (2014) Identification of Quantitative Trait Loci (QTL) Underlying Protein, Oil, and Five Major Fatty Acids' Contents in Soybean. American Journal of Plant Sciences 05:158–167. doi: 10.4236/ajps.2014.51021

Akond M, Liu S, Schoener L, et al (2017) A SNP-Based Genetic Linkage Map of Soybean Using the SoySNP6K Illumina Infinium BeadChip Genotyping Array. Plant Genetics, Genomics, and Biotechnology 1:80–89. doi: 10.5147/pggb.v1i3.154

Akpertey A, Belaffif M, Graef GL, et al (2014) Effects of Selective Genetic Introgression from Wild Soybean to Soybean. Crop Science 54:2683–2695. doi: 10.2135/cropsci2014.03.0189

Alonso-Blanco C, Koornneef M, Stam P (1998) The Use of Recombinant Inbred Lines (RIL) for Genetic Mapping. Arabidopsis Protocols 137–146. doi: 10.1385/0-89603-391-0:137

Andresen JA, Alagarswamy G, Rotz CA, et al (2001) Weather Impacts on Maize, Soybean, and Alfalfa Production in the Great Lakes Region, 1895-1996. Agronomy Journal 93:1059–1070. doi: 10.2134/agronj2001.9351059x

Assefa Y, Purcell LC, Salmeron M, et al (2019) Assessing Variation in US Soybean Seed Composition (Protein and Oil). Frontiers in Plant Science. doi: 10.3389/fpls.2019.00298

Bandillo N, Jarquin D, Song Q, et al (2015) A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. The Plant Genome. doi: 10.3835/plantgenome2015.04.0024

Bazzer SK, Kaler AS, Ray JD, et al (2020) Identification of quantitative trait loci for carbon isotope ratio (δ13C) in a recombinant inbred population of soybean. Theoretical and Applied Genetics 133:2141–2155. doi: 10.1007/s00122-020-03586-0

Beche E, Gillman JD, Song Q, et al (2020) Nested association mapping of important agronomic traits in three interspecific soybean populations. Theoretical and Applied Genetics 133:1039–1054. doi: 10.1007/s00122-019-03529-4

Bernardo R (1994) Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Science 34:20–25. doi: 10.2135/cropsci1994.0011183x003400010003x

Bernardo RN (2020) Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN

Bellaloui N, Smith JR, Ray JD, Gillen AM (2009) Effect of Maturity on Seed Composition in the Early Soybean Production System as Measured on Near-Isogenic Soybean Lines. Crop Science 49:608–620. doi: 10.2135/cropsci2008.04.0192

Boehm JD, Abdel-Haleem H, Schapaugh WT, et al (2019) Genetic Improvement of US Soybean in Maturity Groups V, VI, and VII. Crop Science 59:1838–1852. doi: 10.2135/cropsci2018.10.0627

Boerma HR, Specht JE (2004) Soybeans: improvement, production, and uses. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America

Borevitz JO, Nordborg M (2003) The Impact of Genomics on the Study of Natural Variation in Arabidopsis: Figure 1. Plant Physiology 132:718–725. doi: 10.1104/pp.103.023549

Boudhrioua C, Bastien M, Torkamaneh D, Belzile F (2019) Genome-wide association mapping of Sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. doi: 10.21203/rs.2.14709/v2

Brechenmacher L, Nguyen TH, Zhang N, et al (2015) Identification of Soybean Proteins and Genes Differentially Regulated in Near Isogenic Lines Differing in Resistance to Aphid Infestation. Journal of Proteome Research 14:4137–4146. doi: 10.1021/acs.jproteome.5b00146

Broman KW (2004) The Genomes of Recombinant Inbred Lines. Genetics 169:1133–1146. doi: 10.1534/genetics.104.035212

Broman KW (2011) Guide to qtl mapping with r/qtl. Springer-Verlag New York

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890. doi: 10.1093/bioinformatics/btg112

Brzostowski LF, Diers BW (2017) Agronomic Evaluation of a High Protein Allele from PI407788A on Chromosome 15 across Two Soybean Backgrounds. Crop Science 57:2972–2978. doi: 10.2135/cropsci2017.02.0083

Brzostowski LF, Pruski TI, Specht JE, Diers BW (2017) Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. Theoretical and Applied Genetics 130:2315–2326. doi: 10.1007/s00122-017-2961-x

Buerkle A, Gompert Z (2012) Population genomics based on low coverage sequencing: how low should we go? Molecular Ecology 22:3028–3035. doi: 10.1111/mec.12105

Carter TE, Nelson RL, Sneller CH, Cui Z (2016) Genetic Diversity in Soybean. Agronomy Monographs 303–416. doi: 10.2134/agronmonogr16.3ed.c8

Chen Q-shan, Zhang Z-chen, Liu C-yan, et al (2007) QTL Analysis of Major Agronomic Traits in Soybean. Agricultural Sciences in China 6:399–405. doi: 10.1016/s1671-2927(07)60062-5

Chung J, Babka HL, Graef GL, 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

Collard BC, Jahufer MZ, Brouwer JB, Pang EC (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169–196. doi: 10.1007/s10681-005-1681-5

Concibido V, La Vallee B, Mclaird P, et al (2003) Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. Theoretical and Applied Genetics 106:575–582. doi: 10.1007/s00122-002-1071-5

Dei HK (2011) Soybean as a Feed Ingredient for Livestock and Poultry. Recent Trends for Enhancing the Diversity and Quality of Soybean Products. doi: 10.5772/17601

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm.

Dhanapal AP, Ray JD, Singh SK, et al (2015) Genome-Wide Association Analysis of Diverse Soybean Genotypes Reveals Novel Markers for Nitrogen Traits. The Plant Genome. doi: 10.3835/plantgenome2014.11.0086

Diers BW, Keim P, Fehr WR, Shoemaker RC (1992) RFLP analysis of soybean seed protein and oil content. Theoretical and Applied Genetics 83:608–612. doi: 10.1007/bf00226905

Doerge RW, Zeng Z-B, Weir BS (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statistical Science 12:195–219. doi: 10.1214/ss/1030037909

Doyle, Jj & Doyle, Jl. (1986). A Rapid DNA Isolation Procedure from Small Quantities of Fresh Leaf Tissues. Phytochem Bull. 19.

Eskandari M, Cober ER, Rajcan I (2013) Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theoretical and Applied Genetics 126:1677–1687. doi: 10.1007/s00122-013-2083-z

Evangelista JS, Alves RS, Peixoto MA, et al (2021) Soybean productivity, stability, and adaptability through mixed model methodology. Ciência Rural. doi: 10.1590/0103-8478cr20200406

Falconer DS, Mackay TFC (2009) Introduction to quantitative genetics. Pearson, Prentice Hall, Harlow

Falke KC, Frisch M (2010) Power and false-positive rate in QTL detection with near-isogenic line libraries. Heredity 106:576–584. doi: 10.1038/hdy.2010.87

FAOSTAT, www.fao.org/faostat/en/#search/soybean. Accessed 3/01/2021.

Fasoula VA, Harris DK, Boerma HR (2004) Validation and Designation of Quantitative Trait Loci for Seed Protein, Seed Oil, and Seed Weight from Two Soybean Populations. Crop Science 44:1218–1225. doi: 10.2135/cropsci2004.1218

Fehr WR, Fehr EL, Jessen HJ (1991) Principles of cultivar development. W.R. Fehr, Ames, IA

Fox CM, Cary TR, Colgrove AL, et al (2013) Estimating Soybean Genetic Gain for Yield in the Northern United States-Influence of Cropping History. Crop Science 53:2473–2482. doi: 10.2135/cropsci2012.12.0687

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proceedings of the National Academy of Sciences 97:4718–4723. doi: 10.1073/pnas.97.9.4718

Fu Y-B (2015) Understanding crop genetic diversity under modern plant breeding. Theoretical and Applied Genetics 128:2131–2142. doi: 10.1007/s00122-015-2585-y

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Molecular Ecology 26:5369–5406. doi: 10.1111/mec.14264

Funatsuki H, Suzuki M, Hirose A, et al (2014) Molecular basis of a shattering resistance boosting global dissemination of soybean. Proceedings of the National Academy of Sciences 111:17797–17802. doi: 10.1073/pnas.1417282111

Geladi P, MacDougall D, Martens H (1985) Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Applied Spectroscopy 39:491–500. doi: 10.1366/0003702854248656

Gelli M, Mitchell SE, Liu K, et al (2016) Mapping QTLs and association of differentially expressed gene transcripts for multiple agronomic traits under different nitrogen levels in sorghum. BMC Plant Biology. doi: 10.1186/s12870-015-0696-x

Gillman JD, Tetlow A, Lee J-D, et al (2011) Loss-of-function mutations affecting a specific Glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. BMC Plant Biology 11:155. doi: 10.1186/1471-2229-11-155

Gizlice Z, Carter TE, Burton JW (1993) Genetic Diversity in North American Soybean: I. Multivariate Analysis of Founding Stock and Relation to Coefficient of Parentage. Crop Science 33:614–620. doi: 10.2135/cropsci1993.0011183x003300030038x

Glover KD, Wang D, Arelli PR, et al (2004) Near Isogenic Lines Confirm a Soybean Cyst Nematode Resistance Gene from PI 88788 on Linkage Group J. Crop Science 44:1505–1505. doi: 10.2135/cropsci2004.1505a

Goodstein DM, Shu S, Howson R, et al (2011) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research. doi: 10.1093/nar/gkr944

Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. Trends in Plant Science 15:529–537. doi: 10.1016/j.tplants.2010.05.008

Guo J, Wang Y, Song C, et al (2010) A single origin and moderate bottleneck during domestication of soybean (Glycine max): implications from microsatellites and nucleotide sequences. Annals of Botany 106:505–514. doi: 10.1093/aob/mcq125

Ha B-K, Vuong TD, Velusamy V, et al (2013) Genetic mapping of quantitative trait loci conditioning salt tolerance in wild soybean (Glycine soja) PI 483463. Euphytica 193:79–88. doi: 10.1007/s10681-013-0944-9

Haldane JB, Waddington CH (1931) INBREEDING AND LINKAGE. Genetics 16:504–504. doi: 10.1093/genetics/16.5.504a

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324. doi: 10.1038/hdy.1992.131

Hammond EG, Johnson LA, Su C, et al (2005) Soybean Oil. Bailey's Industrial Oil and Fat Products. doi: 10.1002/047167849x.bio041

Harlan JR, de Wet JM, Price EG (1973) Comparative Evolution of Cereals. Evolution 27:311. doi: 10.2307/2406971

Heim CB, Gillman JD (2016) Genotyping-by-Sequencing-Based Investigation of the Genetic Architecture Responsible for a ~Sevenfold Increase in Soybean Seed Stearic Acid. G3: Genes|Genomes|Genetics 7:299–308. doi: 10.1534/g3.116.035741

Hartwig EE, Kilen TC (1991) Yield and Composition of Soybean Seed from Parents with Different Protein, Similar Yield. Crop Science 31:290–292. doi: 10.2135/cropsci1991.0011183x003100020011x

Huang J, Ma Q, Cai Z, et al (2020) Identification and Mapping of Stable QTLs for Seed Oil and Protein Content in Soybean [Glycine max(L.) Merr.]. Journal of Agricultural and Food Chemistry 68:6448–6460. doi: 10.1021/acs.jafc.0c01271

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

Hymowitz T, Collins FI, Panczner J, Walker WM (1972) Relationship Between the Content of Oil, Protein, and Sugar in Soybean Seed 1. Agronomy Journal 64:613–616. doi: 10.2134/agronj1972.00021962006400050019x

Hyten DL, Pantalone VR, Sams CE, et al (2004) Seed quality QTL in a prominent soybean population. Theoretical and Applied Genetics 109:552–561. doi: 10.1007/s00122-004-1661-5

Hyten DL, Song Q, Zhu Y, et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences 103:16666–16671. doi: 10.1073/pnas.0604379103

Jaganathan D, Bohra A, Thudi M, Varshney RK (2020) Fine mapping and gene cloning in the post-NGS era: advances and prospects. Theoretical and Applied Genetics 133:1791–1810. doi: 10.1007/s00122-020-03560-w

Jander G, Norris SR, Rounsley SD, et al (2002) Arabidopsis Map-Based Cloning in the Post-Genome Era. Plant Physiology 129:440–450. doi: 10.1104/pp.003533

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455. doi: 10.1093/genetics/136.4.1447

Johnson HW, Robinson HF, Comstock RE (1955) Genotypic and Phenotypic Correlations in Soybeans and Their Implications in Selection 1. Agronomy Journal 47:477–483. doi: 10.2134/agronj1955.00021962004700100008x

Joshi T, Fitzpatrick MR, Chen S, et al (2013) Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. Nucleic Acids Research. doi: 10.1093/nar/gkt905

Kabelka EA, Diers BW, Fehr WR, et al (2004) Putative Alleles for Increased Yield from Soybean Plant Introductions. Crop Science 44:784–791. doi: 10.2135/cropsci2004.7840

Kadam S, Vuong TD, Qiu D, et al (2016) Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. Plant Science 242:342–350. doi: 10.1016/j.plantsci.2015.08.015

Keurentjes JJ, Bentsink L, Alonso-Blanco C, et al (2006) Development of a Near-Isogenic Line Population ofArabidopsis thalianaand Comparison of Mapping Power With a Recombinant Inbred Line Population. Genetics 175:891–905. doi: 10.1534/genetics.106.066423

Kim JH, Bae DN, Park S-K, et al (2017) Molecular Genetic Analysis of a Novel Recessive White Flower Gene in Wild Soybean. Crop Science 57:3027–3034. doi: 10.2135/cropsci2017.03.0163

Kim M, Hyten DL, Bent AF, Diers BW (2010) Fine Mapping of the SCN Resistance Locus rhg1-b from PI 88788. The Plant Genome. doi: 10.3835/plantgenome2010.02.0001

Kim M, Schultz S, Nelson RL, Diers BW (2016) Identification and Fine Mapping of a Soybean Seed Protein QTL from PI 407788A on Chromosome 15. Crop Science 56:219–225. doi: 10.2135/cropsci2015.06.0340

Koboldt DC, Steinberg KM, Larson DE, et al (2013) The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell 155:27–38. doi: 10.1016/j.cell.2013.09.006

Kole C (2014) Wild Crop Relatives: Genomic and Breeding Resources Legume Crops and Forages. Springer Berlin

La T, Large E, Taliercio E, 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

La TC, Scaboo A (2018) Characterization of a diverse USDA collection of wild soybean (glycine soja siebold & zucc.) accessions and subsequent mapping for seed composition and agronomic traits in a RIL population. (Doctoral disstertation) Retrieved from https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/66386/research.pdf?sequence=1&isAllowed=y

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199. doi: 10.1093/genetics/121.1.185

Leamy LJ, Zhang H, Li C, et al (2017) A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC Genomics. doi: 10.1186/s12864-016-3397-4

Lee J, Hwang Y-S, Kim ST, et al (2017) Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. Food Chemistry 214:248–258. doi: 10.1016/j.foodchem.2016.07.066

Lee J-D, Yu J-K, Hwang Y-H, et al (2008) Genetic Diversity of Wild Soybean (Glycine soja Sieb. and Zucc.) Accessions from South Korea and Other Countries. Crop Science 48:606–616. doi: 10.2135/cropsci2007.05.0257

Lee SH, Bailey MA, Mian MA, et al (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theoretical and Applied Genetics 93-93:649–657. doi: 10.1007/bf00224058

Lestari P, Van K, Lee J, et al (2013) Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. Frontiers in Plant Science. doi: 10.3389/fpls.2013.00176

Li D, Pfeiffer TW, Cornelius PL (2008) Soybean QTL for Yield and Yield Components Associated with Glycine soja Alleles. Crop Science 48:571–581. doi: 10.2135/cropsci2007.06.0361

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi: 10.1093/bioinformatics/btp324

Li H, Ye G, Wang J (2006) A Modified Algorithm for the Improvement of Composite Interval Mapping. Genetics 175:361–374. doi: 10.1534/genetics.106.066811

Li M-W, Muñoz NB, Wong C-F, et al (2016) QTLs Regulating the Contents of Antioxidants, Phenolics, and Flavonoids in Soybean Seeds Share a Common Genomic Region. Frontiers in Plant Science. doi: 10.3389/fpls.2016.00854

Li Y, Guan R, Liu Z, et al (2008) Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. Theoretical and Applied Genetics 117:857–871. doi: 10.1007/s00122-008-0825-0

Li Y-hui, Zhou G, Ma J, et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology 32:1045–1052. doi: 10.1038/nbt.2979

Li, Xu, Yang, Zhao (2019) Dissecting the Genetic Architecture of Seed Protein and Oil Content in Soybean from the Yangtze and Huaihe River Valleys Using Multi-Locus Genome-Wide Association Studies. International Journal of Molecular Sciences 20:3041. doi: 10.3390/ijms20123041

Liu B, Fujita T, Yan Z-H, et al (2007) QTL Mapping of Domestication-related Traits in Soybean (Glycine max). Annals of Botany 100:1027–1038. doi: 10.1093/aob/mcm149

Liu S, Kandoth PK, Lakhssassi N, et al (2017) The soybean GmSNAP18 gene underlies two types of resistance to soybean cyst nematode. Nature Communications. doi: 10.1038/ncomms14822

Liu Y, Khan SM, Wang J, et al (2016) PGen: large-scale genomic variations analysis workflow and browser in SoyKB. BMC Bioinformatics. doi: 10.1186/s12859-016-1227-y

Liu Z, Li H, Fan X, et al (2016) Phenotypic Characterization and Genetic Dissection of Growth Period Traits in Soybean (Glycine max) Using Association Mapping. PLOS ONE. doi: 10.1371/journal.pone.0158602

Lu W, Wen Z, Li H, et al (2012) Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. Theoretical and Applied Genetics 126:425–433. doi: 10.1007/s00122-012-1990-8

Martínez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoretical and Applied Genetics 85:480–488. doi: 10.1007/bf00222330

Masuda, Tadayoshi & Goldsmith, Peter. (2009). World Soybean Production: Area Harvested, Yield, and Long-Term Projections. International Food and Agribusiness Management Review. 12.

McKenna A, Hanna M, Banks E, 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

Medic J, Atkinson C, Hurburgh CR (2014) Current Knowledge in Soybean Composition. Journal of the American Oil Chemists' Society 91:363–384. doi: 10.1007/s11746-013-2407-9

Mello Filho OL, Sediyama CS, Moreira MA, et al (2004) Grain yield and seed quality of soybean selected for high protein content. Pesquisa Agropecuária Brasileira 39:445–450. doi: 10.1590/s0100-204x2004000500006

Merry R, Butenhoff K, Campbell BW, et al (2019) Identification and Fine-Mapping of a Soybean Quantitative Trait Locus on Chromosome 5 Conferring Tolerance to Iron Deficiency Chlorosis. The Plant Genome 12:190007. doi: 10.3835/plantgenome2019.01.0007

Miao L, Yang S, Zhang K, et al (2019) Natural variation and selection in GmSWEET39 affect soybean seed oil content. New Phytologist 225:1651–1666. doi: 10.1111/nph.16250

Money D, Gardner K, Migicovsky Z, et al (2015) LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3: Genes|Genomes|Genetics 5:2383–2390. doi: 10.1534/g3.115.021667

Morr CV (1981) Nitrogen Conversion Factors for Several Soybean Protein Products. Journal of Food Science 46:1362–1363. doi: 10.1111/j.1365-2621.1981.tb04175.x

Muehlbauer GJ, Specht JE, Thomas-Compton MA, et al (1988) Near-Isogenic Lines—A Potential Resource in the Integration of Conventional and Molecular Marker Linkage Maps. Crop Science 28:729–735. doi: 10.2135/cropsci1988.0011183x002800050002x

Nagasaki M, Yasuda J, Katsuoka F, et al (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nature Communications. doi: 10.1038/ncomms9018

Nascimento D, Polo LR, Lazzari F, et al (2018) Genomic Association between SNP Markers and QTLs for Protein and Oil Content in Grain Weight in Soybean (Glycine max). Journal of Scientific Research and Reports 20:1–13. doi: 10.9734/jsrr/2018/44150

Nawaz MA, Yang SH, Chung G (2018) Wild Soybeans: An Opportunistic Resource for Soybean Improvement. Rediscovery of Landraces as a Resource for the Future. doi: 10.5772/intechopen.74973

Nichols DM, Glover KD, Carlson SR, et al (2006) Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits. Crop Science 46:834–839. doi: 10.2135/cropsci2005.05-0168

Oakeson KF, Wagner JM, Mendenhall M, et al (2017) Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerging Infectious Diseases 23:1441–1445. doi: 10.3201/eid2309.170416

Pantalone VR, Rebetzke GJ, Burton JW, Wilson RF (1997) Genetic regulation of linolenic acid concentration in wild soybean Glycine soja accessions. Journal of the American Oil Chemists' Society 74:159–163. doi: 10.1007/s11746-997-0162-5

Park ST, Kim J (2016) Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. International Neurourology Journal. doi: 10.5213/inj.1632742.371

Panter DM, Allen FL (1995) Using Best Linear Unbiased Predictions to Enhance Breeding for Yield in Soybean: I. Choosing Parents. Crop Science 35:397. doi: 10.2135/cropsci1995.0011183x003500020020x

Pathan SM, Vuong T, Clark K, et al (2013) Genetic Mapping and Confirmation of Quantitative Trait Loci for Seed Protein and Oil Contents and Seed Weight in Soybean. Crop Science 53:765–774. doi: 10.2135/cropsci2012.03.0153

Patil G, Chaudhary J, Vuong TD, et al (2017) Development of SNP Genotyping Assays for Seed Composition Traits in Soybean. International Journal of Plant Genomics 2017:1–12. doi: 10.1155/2017/6572969

Patil G, Do T, Vuong TD, et al (2016) Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. Scientific Reports. doi: 10.1038/srep19199

Patil G, Mian R, Vuong T, 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 TD, Kale S, 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

Pawlowski ML, Vuong TD, Valliyodan B, et al (2019) Whole-genome resequencing identifies quantitative trait loci associated with mycorrhizal colonization of soybean. Theoretical and Applied Genetics 133:409–417. doi: 10.1007/s00122-019-03471-5

Picard toolkit. Broad Institute, GitHub repository: Broad Institute, 2018

Phytozome. http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax). (Accessed March 18, 2021)

Pollard DA (2012) Design and Construction of Recombinant Inbred Lines. Methods in Molecular Biology 31–39. doi: 10.1007/978-1-61779-785-9_3

Pratap A, Das A, Kumar S, Gupta S (2021) Current Perspectives on Introgression Breeding in Food Legumes. Frontiers in Plant Science. doi: 10.3389/fpls.2020.589189

Priolli RH, Carvalho CR, Bajay MM, et al (2019) Genome analysis to identify SNPs associated with oil content and fatty acid components in soybean. Euphytica. doi: 10.1007/s10681-019-2378-5

Qi X, Li M-W, Xie M, et al (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nature Communications. doi: 10.1038/ncomms5340

Qi Z-ming, Wu Q, Han X, et al (2011) Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179:499–514. doi: 10.1007/s10681-011-0386-1

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Ray JD, Dhanapal AP, Singh SK, et al (2015) Genome-Wide Association Study of Ureide Concentration in Diverse Maturity Group IV Soybean [Glycine max (L.) Merr.] Accessions. G3: Genes|Genomes|Genetics 5:2391–2403. doi: 10.1534/g3.115.021774

Rincker K, Nelson R, Specht J, et al (2014) Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. Crop Science 54:1419–1432. doi: 10.2135/cropsci2013.10.0665

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19:491–504. doi: 10.1038/s41576-018-0016-z

Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi: 10.1038/nature08670

Sebolt AM, Shoemaker RC, Diers BW (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

Sedivy EJ, Wu F, Hanzawa Y (2017) Soybean domestication: the origin, genetic architecture and molecular bases. New Phytologist 214:539–553. doi: 10.1111/nph.14418

Seo J-H, Kim K-S, Ko J-M, et al (2018) Quantitative trait locus analysis for soybean (Glycine max) seed protein and oil concentrations using selected breeding populations. Plant Breeding 138:95–104. doi: 10.1111/pbr.12659

Sillanpää MJ, Arjas E (1998) Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data. Genetics 148:1373–1388. doi: 10.1093/genetics/148.3.1373

Silva MA, Muniz AS, Mannigel AR, et al (2011) Monitoring and evaluation of need for nitrogen fertilizer topdressing for maize leaf chlorophyll readings and the relationship with grain yield. Brazilian Archives of Biology and Technology 54:665–674. doi: 10.1590/s1516-89132011000400004

Song J, Liu Z, Hong H, et al (2016) Identification and Validation of Loci Governing Seed Coat Color by Combining Association Mapping and Bulk Segregation Analysis in Soybean. PLOS ONE. doi: 10.1371/journal.pone.0159064

Song Q, Hyten DL, Jia G, et al (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE. doi: 10.1371/journal.pone.0054985

Song Q, Hyten DL, Jia G, et al (2015) Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3: Genes|Genomes|Genetics 5:1999–2006. doi: 10.1534/g3.115.019000

Song Q, Yan L, Quigley C, et al (2020) Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. The Plant Journal 104:800–811. doi: 10.1111/tpj.14960

Spain SL, Barrett JC (2015) Strategies for fine-mapping complex traits. Human Molecular Genetics. doi: 10.1093/hmg/ddv260

Specht JE, Hume DJ, Kumudini SV (1999) Soybean Yield Potential-A Genetic and Physiological Perspective. Crop Science 39:1560–1570. doi: 10.2135/cropsci1999.3961560x

Spielbauer G, Armstrong P, Baier JW, et al (2009) High-Throughput Near-Infrared Reflectance Spectroscopy for Predicting Quantitative and Qualitative Composition Phenotypes of Individual Maize Kernels. Cereal Chemistry Journal 86:556–564. doi: 10.1094/cchem-86-5-0556

Stein HH, Berger LL, Drackley JK, et al (2008) Nutritional Properties and Feeding Values of Soybeans and Their Coproducts. Soybeans 613–660. doi: 10.1016/b978-1-893997-64-6.50021-4

Stupar RM (2010) Into the wild: The soybean genome meets its undomesticated relative. Proceedings of the National Academy of Sciences 107:21947–21948. doi: 10.1073/pnas.1016809108

Sundaramoorthy J, Park GT, Chang JH, et al (2016) Identification and Molecular Analysis of Four New Alleles at the W1 Locus Associated with Flower Color in Soybean. PLOS ONE. doi: 10.1371/journal.pone.0159865

USB, 2019. United Soybean Board Supply & Disappearance. USB Market View Database (n.d.). Available at: https://marketviewdb.centrec.com/sd/. (Accessed: March 3, 2021)

Uses for Soybeans | Missouri Soybean. (2012). Retrieved December 27, 2019, from Mosoy.org website: https://mosoy.org/check-off-at-work/domestic-marketing/

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of Quantitative Trait Loci for Protein and Lipid Contents in Soybean Seeds Using Recombinant Inbred Lines. Breeding Science 53:133–140. doi: 10.1270/jsbbs.53.133

Uga Y, Sugimoto K, Ogawa S, et al (2013) Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. Nature Genetics 45:1097–1102. doi: 10.1038/ng.2725

USB, 2019. USB Market View Database Legacy, marketviewdb.centrec.com/sd/. Accessed 3/01/2021.

USDA Soybean Germplasm Collection. In: GBIF. https://www.gbif.org/grscicoll/collection/6e5b27ae-183f-47c1-8a60-7dda5fe05b11. Accessed 10/12/2020.

Van K, McHale L (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 18:1180. doi: 10.3390/ijms18061180

Von Korff M, Wang H, Léon J, Pillen K (2004) Development of candidate introgression lines using an exotic barley accession (Hordeum vulgare ssp. spontaneum) as donor. Theoretical and Applied Genetics 109:1736–1745. doi: 10.1007/s00122-004-1818-2

Wang J, Chen P, Wang D, et al (2015) Identification and mapping of stable QTL for protein content in soybean seeds. Molecular Breeding. doi: 10.1007/s11032-015-0285-6

Wang K-J, Takahata Y (2007) A Preliminary Comparative Evaluation of Genetic Diversity between Chinese and Japanese Wild Soybean (Glycine soja) Germplasm Pools using SSR markers. Genetic Resources and Crop Evolution 54:157–165. doi: 10.1007/s10722-005-2641-6

Wang P-wu, Di Q, Liu X-Y (2020) Genome-Wide association Study Identifies Candidate Genes Related to Oleic acid content of Soybean Seed. doi: 10.21203/rs.3.rs-17853/v1

Warrington CV, Abdel-Haleem H, Hyten DL, 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

Watanabe S, Xia Z, Hideshima R, et al (2011) A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals that theGIGANTEAGene Is Involved in Soybean Maturity and Flowering. Genetics 188:395–407. doi: 10.1534/genetics.110.125062

Whittaker JC, Thompson R, Visscher PM (1996) On the mapping of QTL by regression of phenotype on marker-type. Heredity 77:23–32. doi: 10.1038/hdy.1996.104

Wilson, R.F. 2004. Seed composition. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, Production, and Uses. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.: 621-677

Xia Z, Watanabe S, Yamada T, et al (2012) Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proceedings of the National Academy of Sciences. doi: 10.1073/pnas.1117982109

Ye H, Song L, Chen H, et al (2018) A major natural genetic variation associated with root system architecture and plasticity improves waterlogging tolerance and yield in soybean. Plant, Cell & Environment. doi: 10.1111/pce.13190

Young ND, Zamir D, Ganal MW, Tanksley SD (1988) Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the tm-2a gene in tomato. Genetics 120:579–585. doi: 10.1093/genetics/120.2.579

Yu X, Yuan F, Fu X, Zhu D (2016) Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. Food Chemistry 196:776–782. doi: 10.1016/j.foodchem.2015.09.092

Yuan G, Wan Y, Li X, et al (2017) Development of Near-Isogenic Lines in a Parthenogenetically Reproduced Thrips Species, Frankliniella occidentalis. Frontiers in Physiology. doi: 10.3389/fphys.2017.00130

Zeng S, Lyu Z, Narisetti SR, et al (2019) Knowledge Base Commons (KBCommons) v1.1: a universal framework for multi-omics data integration and biological discoveries. BMC Genomics. doi: 10.1186/s12864-019-6287-8

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468. doi: 10.1093/genetics/136.4.1457

Zhang D, Cheng H, Hu Z, et al (2012) Fine mapping of a major flowering time QTL on soybean chromosome 6 combining linkage and association analysis. Euphytica 191:23–33. doi: 10.1007/s10681-012-0840-8

Zhang H, Song Q, Griffin JD, Song B-H (2017) Genetic architecture of wild soybean (Glycine soja) response to soybean cyst nematode (Heterodera glycines). Molecular Genetics and Genomics 292:1257–1265. doi: 10.1007/s00438-017-1345-x

Zhang J, Wang X, Lu Y, et al (2018) Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. Molecular Plant 11:460–472. doi: 10.1016/j.molp.2017.12.016

Zhang T, Wu T, Wang L, et al (2019) A Combined Linkage and GWAS Analysis Identifies QTLs Linked to Soybean Seed Protein and Oil Content. International Journal of Molecular Sciences 20:5915. doi: 10.3390/ijms20235915

Zhang W-K, Wang Y-J, Luo G-Z, et al (2004) QTL mapping of ten agronomic traits on the soybean (Glycine max L. Merr.) genetic map and their association with EST markers. Theoretical and Applied Genetics 108:1131–1139. doi: 10.1007/s00122-003-1527-2

Zhang YH, Liu MF, He JB, et al (2015) Marker-assisted breeding for transgressive seed protein content in soybean [Glycine max (L.) Merr.]. Theoretical and Applied Genetics 128:1061–1072. doi: 10.1007/s00122-015-2490-4

Zhong C, Sun S, Zhang X, et al (2020) Fine Mapping, Candidate Gene Identification and Co-segregating Marker Development for the Phytophthora Root Rot Resistance Gene RpsYD25. Frontiers in Genetics. doi: 10.3389/fgene.2020.00799

Zhou Z, Jiang Y, Wang Z, 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

Zhu X, Leiser WL, Hahn V, Würschum T (2021) Identification of seed protein and oil related QTL in 944 RILs from a diallel of early-maturing European soybean. The Crop Journal 9:238–247. doi: 10.1016/j.cj.2020.06.006

Table 2-1. Descriptive statistic of minimum, maximum, means, ranges, standard deviation (SD), coefficient of variation (CV), skewness, and kurtosis of seed oil and protein, and least square means of seed oil and protein between environments.

traits	environment[a]	min[d]	max[d]	mean[d]	SD[e]	CV (%)[f]	skewness	kurtosis	groups[g]
oil	18/19GH	20.1	22.1	21.1	0.51	2.41	0.08	-0.80	a
	19CLM	17.8	19.8	18.9	0.45	2.41	-0.05	-0.27	b
	19NOV	17.8	20.3	19.0	0.59	3.10	0.16	-0.39	b
	20CLM	16.2	17.9	17.2	0.43	2.49	-0.32	-0.48	c
	20NOV	16.2	18.3	17.1	0.48	2.83	0.27	-0.45	c
	CLM&NOV[b]	17.4	18.9	18.0	0.36	1.98	0.18	-0.51	
	Combined[c]	18.0	19.5	18.6	0.34	1.82	0.27	-0.14	
protein	18/19GH	38.2	44.8	41.4	1.47	3.55	-0.10	-0.12	d
	19CLM	42.8	45.7	44.1	0.83	1.87	0.23	-1.03	a
	19NOV	41.4	46.6	43.8	0.96	2.20	0.50	1.31	bc
	20CLM	42.9	45.2	44.0	0.60	1.36	-0.11	-0.80	ab
	20NOV	42.7	45.4	43.8	0.73	1.66	0.46	-0.53	c
	CLM&NOV[b]	42.8	45.2	44.0	0.60	1.36	0.23	-0.83	
	Combined[c]	43.0	44.9	43.4	0.69	1.59	0.11	-0.81	

[a]Five environments (2018/2019 Greenhouse, 2019 in Columbia, 2019 in Novelty, 2020 in Columbia, and 2020 in Novelty) were assigned as 18/19GH, 19CLM, 19NOV, 20CLM, and 20NOV; [b]Combined seed oil and protein content from four field environments (19CLM, 19NOV, 20CLM, and 20NOV); [c]Combined seed oil and protein content from five environments (18/19GH, 19CLM, 19NOV, 20CLM, and 20NOV); [d]Standard Deviation; [e]Coefficient of Variation; [f]Groupings of least square means

Table 2-2. Pearson Correlation between seed oil and protein in the high protein RHD-NIL population across multiple environments.

		18/19GH		19CLM		19NOV		20CLM		20NOV		CLM&NOV[a]		Combined[b]	
Environment	Trait	Oil	Protein	Oil	Protein	Oil	Protein	Oil	Protein	Oil	Protein	Oil	Protein	Oil	Protein
18/19GH	Oil	1													
	Protein	-0.75***	1												
19CLM	Oil	0.46**	-0.49**	1											
	Protein	-0.48**	0.45**	-0.77***	1										
19NOV	Oil	0.27[ns]	-0.38*	0.29[ns]	-0.29ns	1									
	Protein	-0.41**	0.48**	-0.28ns	0.38*	-0.78***	1								
20CLM	Oil	0.42**	-0.38*	0.55***	-0.39*	0.32*	-0.31*	1							
	Protein	-0.40**	0.54***	-0.53***	0.53***	-0.27ns	0.39*	-0.55***	1						
20NOV	Oil	0.57***	-0.58***	0.54***	-0.50**	0.25[ns]	-0.37*	0.39*	-0.70***	1					
	Protein	-0.51**	0.49**	-0.54***	0.60***	-0.22ns	0.36*	-0.34*	0.55***	-0.70***	1				
CLM&NOV[a]	Oil	0.58***	-0.62***	0.79***	-0.65***	0.68***	-0.63***	0.74***	-0.68***	0.73***	-0.60***	1			
	Protein	-0.59***	0.63***	-0.68***	0.81***	-0.55***	0.73***	-0.50**	0.75***	-0.70***	0.79***	-0.83***	1		
Combined[b]	Oil	0.74***	-0.66***	0.79***	-0.65***	0.51**	-0.52***	0.76***	-0.67***	0.74***	-0.60***	0.94***	-0.78***	1	
	Protein	-0.72***	0.83***	-0.69***	0.75***	-0.50**	0.66***	-0.52***	0.76***	-0.74***	0.76***	-0.83***	0.94***	-0.85***	1

[a]Combined seed oil and protein content from four field environments (19CLM, 19NOV, 20CLM, and 20NOV)
[b]Combined seed oil and protein content from five field environments (18/19GH, 19CLM, 19NOV, 20CLM, and 20NOV)
*Indicates significant at the 0.05 level ($P < 0.05$)
**Indicates significant at the 0.01 level ($P < 0.01$)
***Indicates significant at the 0.001 level ($P < 0.001$)
[ns]Indicates not significant

Table 2-3. Summary of the analysis of variance for seed protein and seed oil with heritability (h^2) on an entry-mean basis.

Source of Variance	Df	Mean Sq	F-value	Pr(>F)	Mean Sq	F-value	Pr(>F)
			---------- Protein ----------		---------- Oil ----------		
Genotype (G)	49	2.27	4.91	9.99E-15***	0.81	4.21	3.23E-12***
Environment (E)	3	1.86	4.03	8.46E-03**	92.68	480.91	<2.22E15***
Genotype x Environment (GxE)	139	0.64	1.38	2.40E-02*	0.25	1.28	6.61E-02
Replications in Environment	4	1.66	3.61	7.15E-03**	0.07	0.37	8.29E-01
Residual	161	0.46			0.19		
h^2		0.72			0.69		

*Indicates significant at the 0.05 level ($P < 0.05$)
**Indicates significant at the 0.01 level ($P < 0.01$)
***Indicates significant at the 0.001 level ($P < 0.001$)

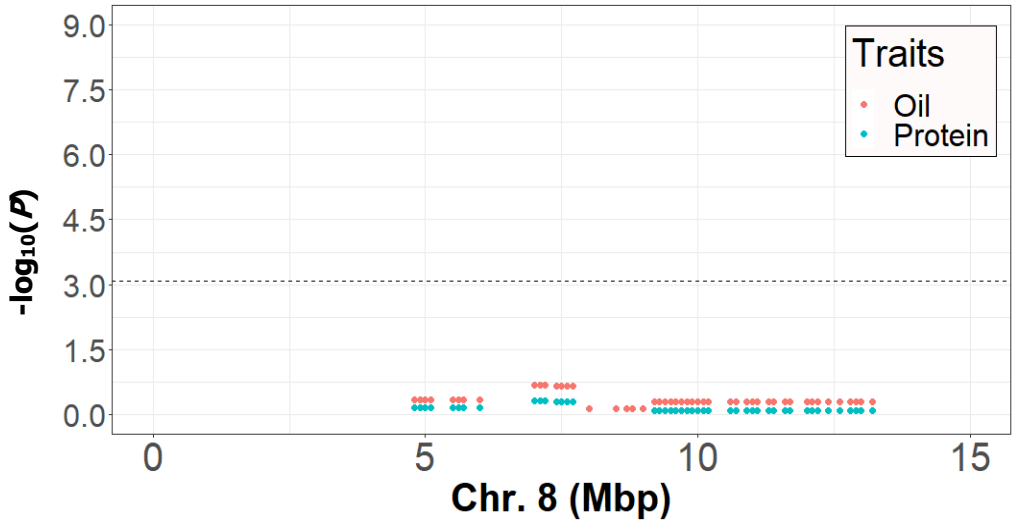

Figure 2-1. Validation of the Chr. 8 oil QTL. 62 SoySNP6K markers -$\log_{10}(P)$ values were plotted across the initial RIL QTL for oil and protein. Significant threshold was at 3.09 -$\log_{10}(P)$ based on a Bonferroni Correction.

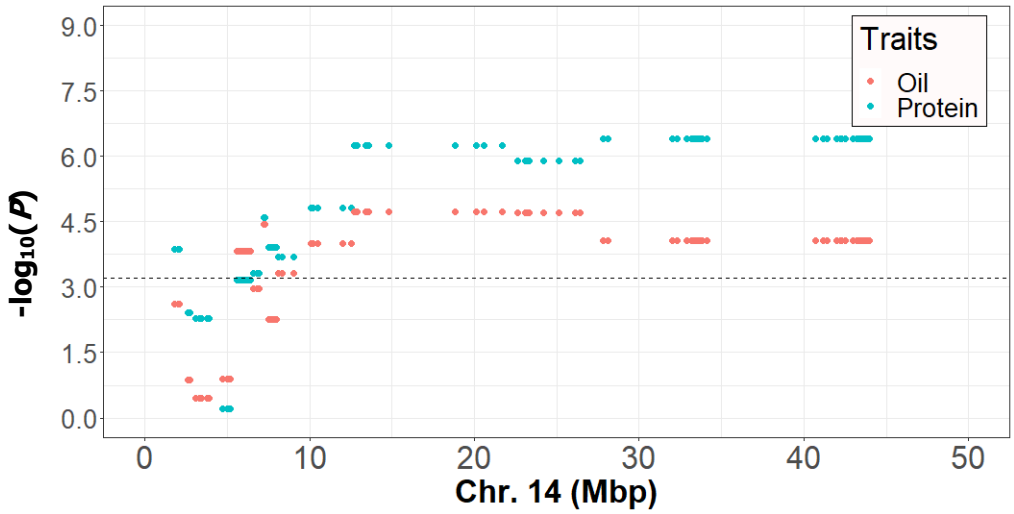

Figure 2-2. Validation of the Chr. 14 oil QTL. 93 SoySNP6K markers -$\log_{10}(P)$ values were plotted across the 20 chromosomes for oil and protein. Significant threshold was at 3.27 -$\log_{10}(P)$ based on a Bonferroni Correction.

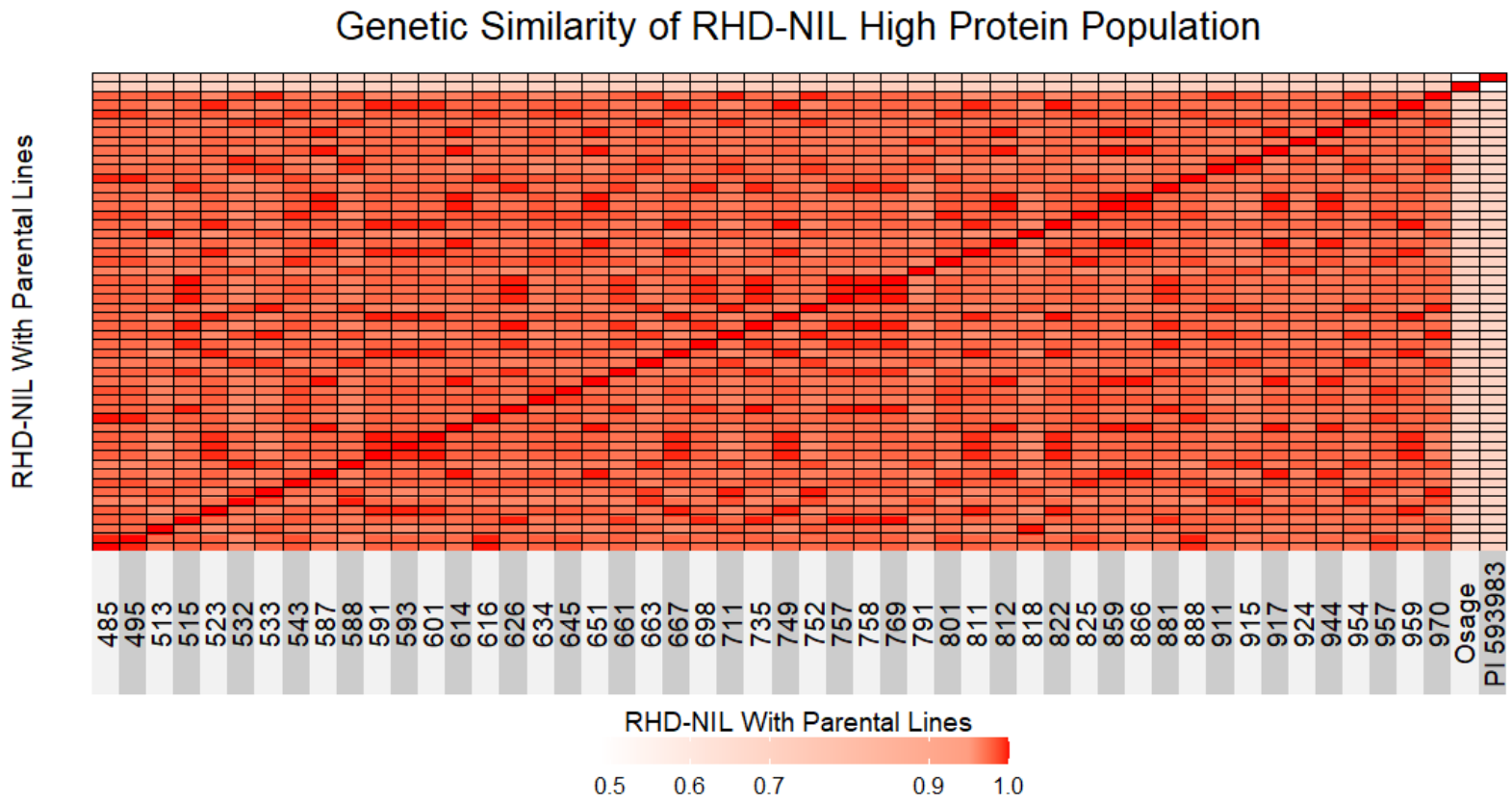

Figure 2-3. Genetic similarity test between individual RHD-NIL and parental lines shown as a heatmap. Red indicates 1.0 genetically similar, light red indicates 0.90 genetically similar, and light pink indicates less than 0.50 genetically similar. Osage represents parent one and PI 593983 represents parent two.

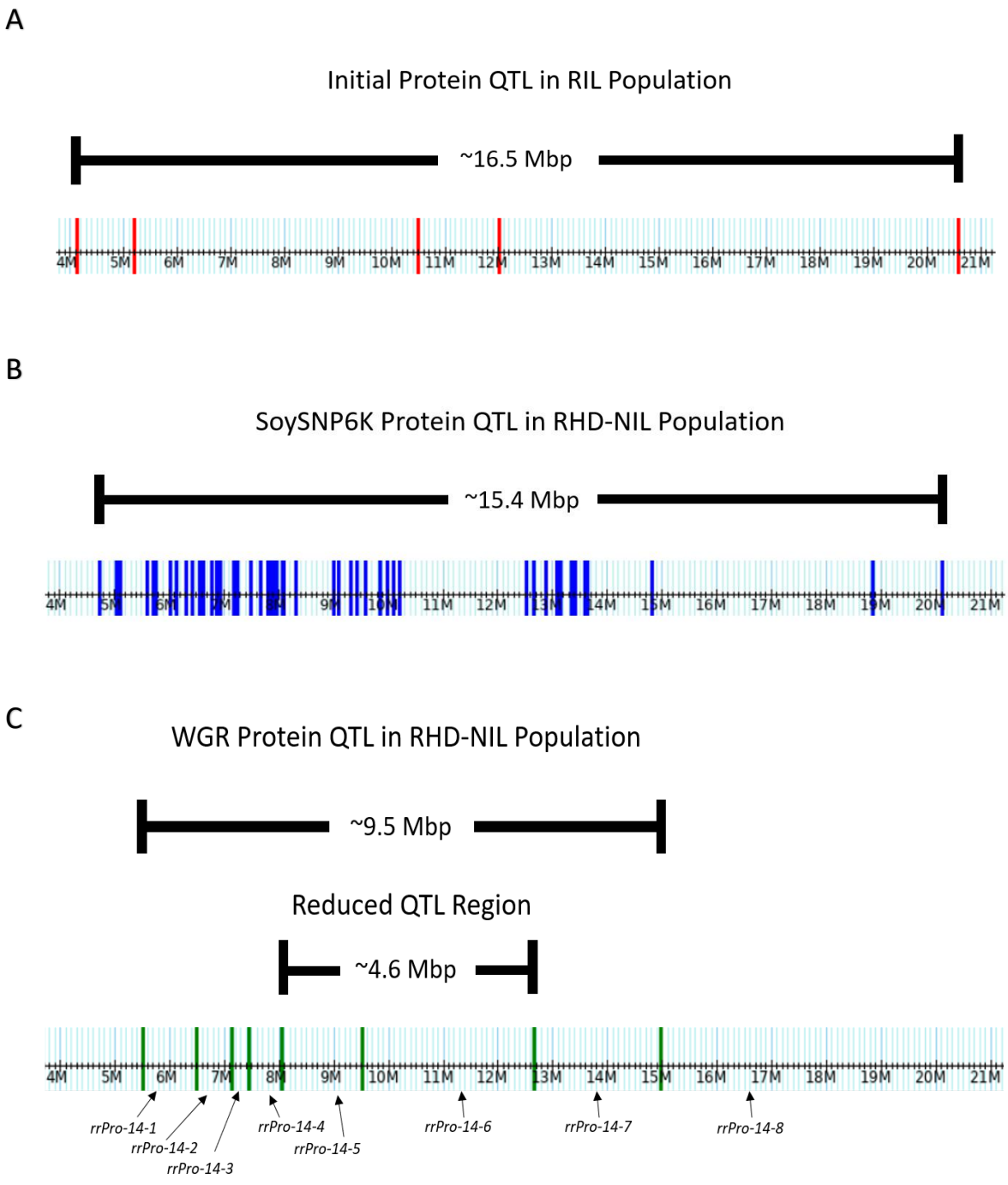

Figure 2-4. Distribution of markers across the Chr.14 protein QTL on the physical map. A) Five genotyping-by-sequencing (GBS) markers in the initial RIL population. B) Fifty-one SoySNP6K markers in the RHD-NIL population. C) Eight WGR markers in the RHD-NIL population. The eight recombination regions are indicted on the physical map.

Table 2-4. The eight recombination regions for seed protein and oil on Chr. 14.

trait	QTL name[a]	Chr.[b]	Recomb. Region[c]	marker interval[d]	position (cM)	R^2 (%)[e]	F-value	Number of Genes[f]
Protein content	*q-14*	14	*rr-14-1*	Gm14_5509372-Gm14_6485179	0.00	10.47	0.39	103
			rr-14-2	Gm14_6487608-Gm14_7138691	3.32	13.13	0.89	53
			rr-14-3	Gm14_7141628-Gm14_7453099	5.46	16.43	3.73*	23
			rr-14-4	Gm14_7455192-Gm14_8048870	16.06	14.65	0.07	44
			rr-14-5	Gm14_8059955-Gm14_9506311	23.41	12.61	5.60**	100
			rr-14-6	Gm14_9508613-Gm14_12648760	25.57	16.16	7.03***	123
			rr-14-7	Gm14_12655776-Gm14_14976378	36.20	17.99	2.02	61
			rr-14-8	Gm14_14976378-Gm14_44140803	37.26	16.75	0.49	1,071
Oil content	*q-14*	14	*rr-14-1*	Gm14_5509372-Gm14_6485179	0.00	3.56	2.77*	103
			rr-14-2	Gm14_6487608-Gm14_7138691	3.32	2.89	0.21	53
			rr-14-3	Gm14_7141628-Gm14_7453099	5.46	2.94	0.47	23
			rr-14-4	Gm14_7455192-Gm14_8048870	16.06	2.82	0.51	44
			rr-14-5	Gm14_8059955-Gm14_9506311	23.41	3.60	1.24	100
			rr-14-6	Gm14_9508613-Gm14_12648760	25.57	4.51	1.93	123
			rr-14-7	Gm14_12655776-Gm14_14976378	36.20	4.81	2.57	61
			rr-14-8	Gm14_14976378-Gm14_44140803	37.26	4.23	0.70	1,071

[a] QTL name for protein and oil on Chr; 14; [b]Chromosome number; [c]Name of recombination regions for protein and oil; [d]Marker interval of the recombination regions; Gm14 represents Chr; 14 and the follow number represents the physical position; [e]Variation explained for protien and oil (R^2) in percentage; [f]Number of total genes present in the recombination regions

*Indicates significant at the 0.1 level ($P < 0.1$)

**Indicates significant at the 0.05 level ($P < 0.05$)

***Indicates significant at the 0.05 level ($P < 0.01$)

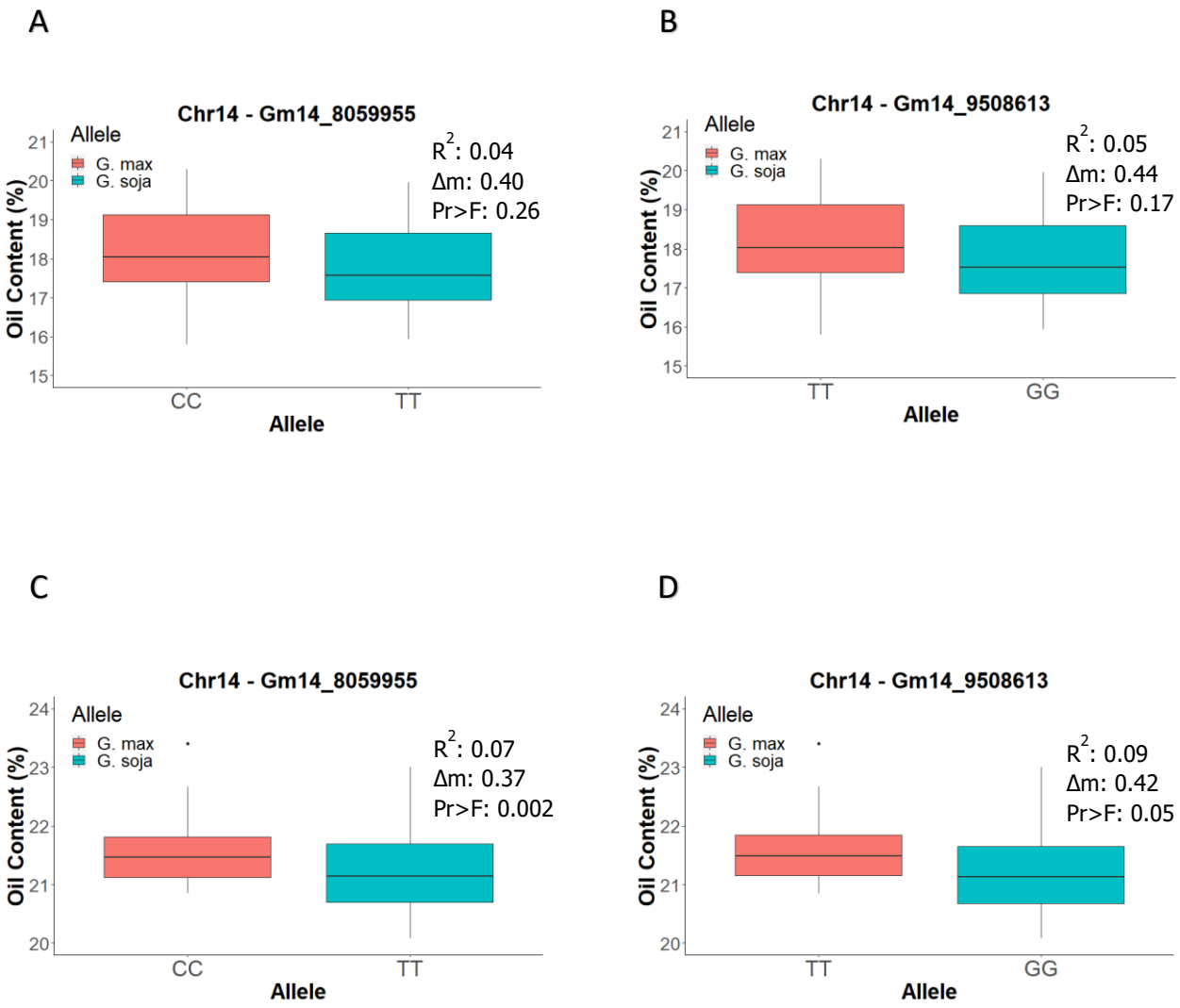

Figure 2-5. The differences in phenotypic values of oil content (%) carrying different homozygous alleles for the markers Gm14_8059955 and Gm14_9508613. Gm14_8059955 represents the recombinant region *rr*-14-5 and is the first marker in *rr*-14-5 (A, C). Gm14_9508613 represents the recombinant region *rr*-14-6 and is the first marker in *rr*-14-6 (B, D). Allele (CC) is the allele from *G. max* (Osage) and (TT) is the allele from *G. soja* (PI 593983) in *rr*-14-5. The alleles in *rr*-14-6 is (TT) for *G. max* (Osage and (GG) for *G. soja* (PI 593983). A) Oil content from CLM&NOV for *rr*-14-5. B) Oil content from 18/19GH for *rr*-14-5. C) Oil content from CLM&NOV for *rr*-14-6. D) Oil content from 18/19GH for *rr*-14-6.

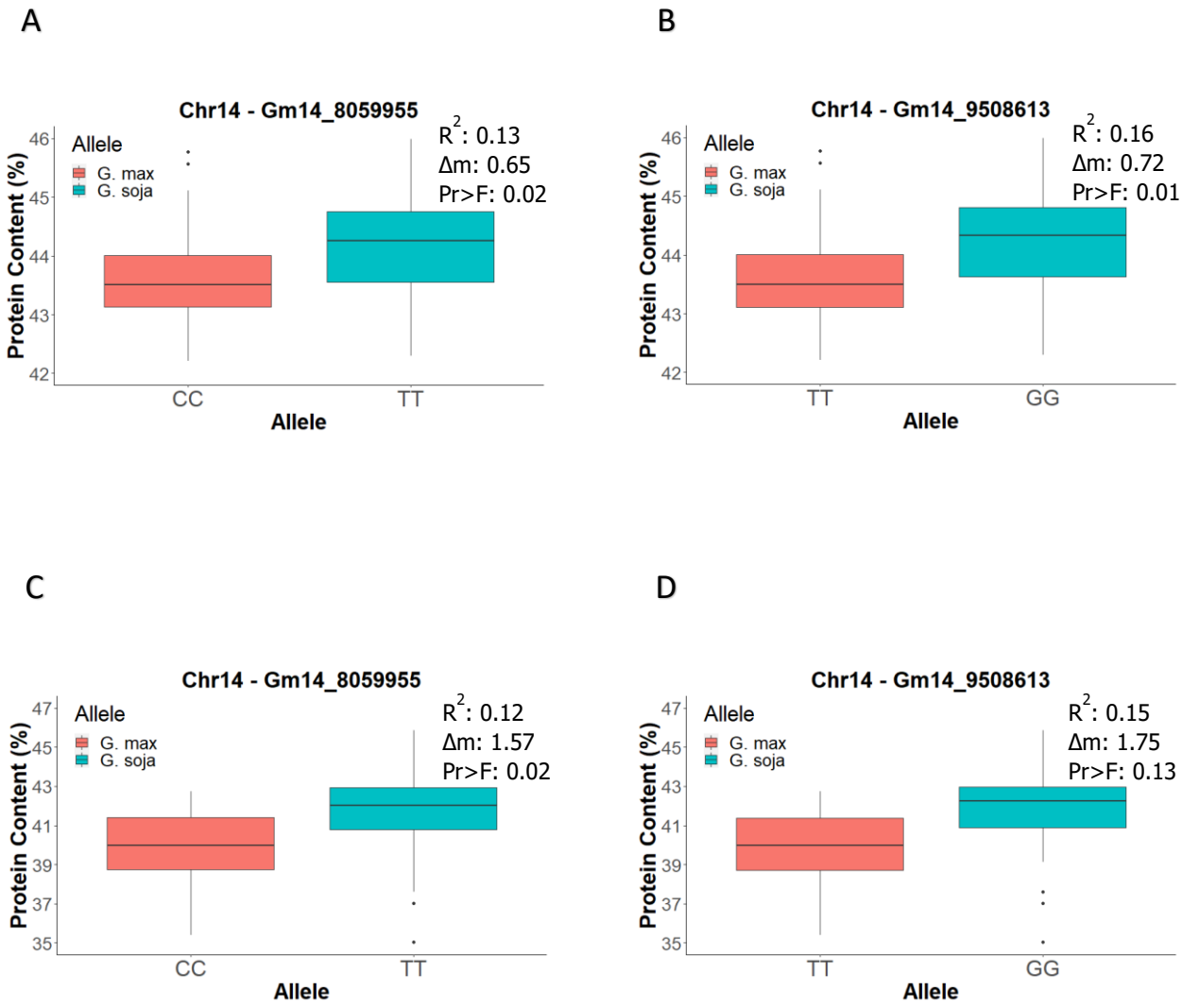

Figure 2-6. The differences in phenotypic values of protein content (%) carrying different homozygous alleles for the markers Gm14_8059955 and Gm14_9508613. Gm14_8059955 represents the recombinant region *rr*-14-5 and is the first marker in *rr*-14-5 (A, C). Gm14_9508613 represents the recombinant region *rr*-14-6 and is the first marker in *rr*-14-6 (B, D). Allele (CC) is the allele from *G. max* (Osage) and (TT) is the allele from *G. soja* (PI 593983) in *rr*-14-5. The alleles in *rr*-14-6 is (TT) for *G. max* Osage and (GG) for *G. soja* (PI 593983). A) Protein content from CLM&NOV for *rr*-14-5. B) Protein content from 18/19GH for *rr*-14-5. C) Protein content from CLM&NOV for *rr*-14-6. D) Protein content from 18/19GH for *rr*-14-6.

Table 2-5. Candidate protein related genes within *rr*-14-5.

Gmax 2.0 Gene IDs	Start	Stop	Biological Process Descriptions	KOG Annotations
Glyma.14G090200	8218662	8222883	amino acid transport	Amino acid transporter protein
Glyma.14G090500	8255191	8260680	protein folding	NA
Glyma.14G090800	8288524	8291197	embryo development ending in seed dormancy	NA
Glyma.14G091000	8354011	8356663	protein acetylation	NA
Glyma.14G091600	8490761	8497722	intracellular protein transport; protein targeting to vacuole	Equilibrative nucleoside transporter protein
Glyma.14G091700	8498431	8505858	protein targeting to vacuole	NA
Glyma.14G092800	8614957	8622417	vacuolar protein processing	Asparaginyl peptidases
Glyma.14G093000	8635762	8637861	N-terminal protein myristoylation	NA
Glyma.14G093300	8671736	8674095	protein glycosylation; ubiquitin-dependent protein catabolic process	NADH-ubiquinone oxidoreductase, NUFS7/PSST/20 kDa subunit
Glyma.14G093400	8675773	8681315	protein glycosylation ubiquitin-dependent protein catabolic process	NADH-ubiquinone oxidoreductase, NUFS7/PSST/20 kDa subunit
Glyma.14G094400	8775073	8785330	protein phosphorylation	Tyrosine kinase specific for activated (GTP-bound) p21cdc42Hs
Glyma.14G094700	8801566	8808263	glycine metabolic process	Glycine dehydrogenase (decarboxylating)

Glyma.14G095200	8850684	8852638	protein targeting to mitochondrion	Small Nuclear ribonucleoprotein splicing factor
Glyma.14G095700	8906064	8920433	protein catabolic process; proteolysis	AAA+-type ATPase
Glyma.14G096000	8960003	8967096	protein catabolic process; proteolysis	AAA+-type ATPase containing the peptidase M41 domain
Glyma.14G096100	8983044	8990246	protein catabolic process; proteolysis	AAA+-type ATPase containing the peptidase M41 domain
Glyma.14G096200	9006426	9009129	amino acid transport	NA
Glyma.14G096600	9045982	9049959	amino acid transport	Beta-fructofuranosidase (invertase)
Glyma.14G096800	9106492	9110189	cellular response to unfolded protein	Heat shock transcription factor
Glyma.14G096900	9125256	9128024	protein targeting to membrane	Cytochrome P450 CYP3/CYP5/CYP6/CYP9 subfamilies
Glyma.14G097200	9161278	9165065	lysine biosynthesis via diaminopimelate	NA
Glyma.14G098000	9247150	9255896	G-protein coupled receptor signaling pathway	G-protein alpha subunit (small G protein superfamily)
Glyma.14G098100	9259148	9266365	cellular modified amino acid biosynthesis	NA
Glyma.14G098900	9438996	9446823	auxin homeostasis; ubiquitin-dependent protein catabolic process	F-box protein containing LRR

Table 2-6. Candidate protein related genes within *rr*-14-6.

Gmax 2.0 Gene IDs	Start	Stop	Biological Process Descriptions	KOG Annotations
Glyma.14G099100	9557728	9561677	protein import into peroxisome matrix; protein monoubiquitination	Predicted E3 ubiquitin ligase, integral peroxisomal membrane protein
Glyma.14G099900	9736779	9737754	protein folding	Molecular chaperone (small heat-shock protein Hsp26/Hsp42)
Glyma.14G100000	9740301	9741067	protein folding	Molecular chaperone (small heat-shock protein Hsp26/Hsp42)
Glyma.14G100300	9818448	9819839	protein phosphorylation	Serine/threonine protein kinase
Glyma.14G100400	9839656	9841977	protein phosphorylation	Serine/threonine protein kinase
Glyma.14G100600	9884437	9887016	protein phosphorylation	Serine/threonine protein kinase
Glyma.14G100700	9887705	9890065	protein phosphorylation	Serine/threonine protein kinase
Glyma.14G100800	9954693	9961410	protein phosphorylation	Serine/threonine protein kinase
Glyma.14G101400	10034501	10040535	N-terminal protein myristoylation; protein dephosphorylation	Protein phosphatase 1B (formerly 2C)
Glyma.14G101500	10046353	10051180	protein maturation; tRNA aminoacylation for protein translation	Aspartyl-tRNA synthetase

Glyma.14G102000	10082031	10086953	N-terminal protein myristoylation	Endosomal membrane proteins, EMP70
Glyma.14G102700	10194805	10198050	aromatic amino acid family biosynthesis	Chorismate mutase
Glyma.14G104800	10748114	10751676	regulation of amino acid import	NA
Glyma.14G104900	10755592	10760847	protein dephosphorylation	Serine/threonine protein phosphatase
Glyma.14G105000	10783830	10786799	protein phosphorylation	NA
Glyma.14G105200	10798891	10799849	regulation of amino acid export	NA
Glyma.14G105700	10891592	10898958	ubiquitin-dependent protein catabolic process	Ubiquitin carboxyl-terminal hydrolase
Glyma.14G105800	10911008	10913393	ubiquitin-dependent protein catabolic process	NA
Glyma.14G105900	10916033	10919283	amino acid transport	NA
Glyma.14G106100	10934325	10935581	N-terminal protein myristoylation	NA
Glyma.14G106900	11322552	11324154	N-terminal protein myristoylation; protein phosphorylation	NA
Glyma.14G107600	11515939	11529083	protein phosphorylation	Serine/threonine protein kinase

Glyma.14G108200	11824403	11825059	N-terminal protein myristoylation	NA
Glyma.14G108400	11901726	11906954	protein dephosphorylation; regulation of seed germination	Serine/threonine protein phosphatase
Glyma.14G109000	12020173	12024201	ubiquitin-dependent protein catabolic process	Ubiquitin-specific protease
Glyma.14G111000	12575744	12578831	proteolysis; ubiquitin-dependent protein catabolic process	NA

Supplementary Table 2-1. Ground soybean NIRS calibrations for 2018, 2019, and 2020.

Parameter	2018 Oil	2018 Protein	2019 Oil	2019 Protein	2020 Oil	2020 Protein
Calibration Type	Honigs Regression	Honigs Regression	Honigs Regression	Honigs Regression	Honigs Regression	Honigs Regression
Moisture basis	Dry basis	Dry basis	Dry basis	Dry basis	Dry basis	Dry basis
Parameter Unit	%	%	%	%	%	%
SECV[a]	0.84	0.86	0.89	0.91	0.71	0.75
R^2CV[b]	0.93	0.94	0.79	0.86	0.86	0.90
Min[c]	9.92	33.44	13.56	33.39	13.56	32.62
Max[c]	26.47	53.55	26.47	53.55	26.47	53.55
n[d]	3537	3790	3884	4137	5083	6591
Date[e]	2018	2018	6/16/2019	6/16/2019	7/16/2020	7/16/2020

[a]Standard error of cross validation; [b]Coefficient of determination of cross validation; [c]Minimum and maximum value; [d]Number of samples; [e]Last updated date

Chapter III

Linkage Analysis for Whole Plant Biomass, Carbon, Nitrogen and Seed Composition using a RIL Mapping Population

Abstract

The demand for soybean has increased due to the seed protein and oil content and for its many uses. Soybean seeds requires a large amount of nitrogen because of its high seed protein content. Through a symbiotic association between soil microorganisms and soybean root nodules, soybean is able to fix atmospheric dinitrogen for nitrogen uptake. In our study, we conducted linkage analysis for a recombinant inbred line (RIL) mapping population for plant biomass content, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content. Plant biomass was collected by bulking five soybean shoot samples per plot from 261 plots in one location and 262 plots from three locations and bulking three soybean shoots samples per plot from 262 plots in one location. Plant materials were dried and weighed for whole plant biomass weight. Whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content was analyzed via near infrared spectroscopy. The objective of this study was to examine nitrogen mobilization from a mapping population from the cross PI 361103 (contains high shoot N content and low seed N content) x PI 567572B (contains high seed N content and low shoot N content), identify QTL for plant biomass, whole plant carbon content, whole plant nitrogen content, and seed composition, and study maternal effects of cytoplasmic inheritance of the five traits from the reciprocal parental cross. Linkage analysis was conducted using BARCSoySNP50K markers. We identified six QTL for plant biomass, two QTL for whole plant carbon content, three QTL for whole plant nitrogen content, three QTL for seed oil content, and five QTL for seed protein content, with multiple traits having overlapping

QTL intervals. Our results indicate QTL associated with multiple traits demonstrating the potential of pleiotropic effect in our mapping population.

Introduction

Soybean [*Glycine max* (L.) Merr] is one of the most valuable crops in the world due to the high protein and oil of its seed, which have uses as feed for livestock, a good source of protein and oil for human health, and as a biofuel stock (Masuda et al., 2009). The world total soybean production in 2019 was approximately 334 million metric tons (FAOSTAT, 2020). Soybean seed contain approximately 20% oil and 40% protein on a dry weight basis, and the two most important economical components of soybean are its oil and soybean meal (Wilson, 2004; Warrington et al., 2015). The increased use of soybean meal in animal feed as a protein source has been a major driving force in soybean production (Dei, 2011).

Soybean can fix atmospheric dinitrogen (N_2) within root nodules through symbiotic association with soil microorganisms, typically of the genus *Bradyrhizobia* (Bohlool et al., 1992; Keyser et al., 1992). There are three sources by which soybean plants can assimilate nitrogen: 1) symbiotic biological N_2 fixation by root nodules (biological nitrogen fixation, BNF), 2) absorbed from soil mineralized nitrogen, and 3) nitrogen derived from fertilizer application (Harper, 1974; Harper, 1987, Ohyama et al., 2017). Soybean is estimated to require an average of 80 kg nitrogen in aboveground dry matter (ADM) per metric ton to produce seeds (Salvagiotti et al., 2008; Tamagno et al., 2017). The homeostatic relationship between BNF and nitrogen absorbed from the soil are well known; BNF decreases as soybeans uptake more nitrogen from the soil (Santachiara et al., 2017; Streeter & Wong, 1988).

Soybean seeds contain a high concentration of protein which requires a large input of nitrogen (Sinclair and De Wit, 1976; Giller and Cadisch, 1995; Ohyama et al., 2017). Roots uptake nitrogen in the forms of amino acids, ureides, and other nitrogen compounds and are transported to shoot via the xylem (Rentsch et al., 2007). Plant leaves and roots are the largest sinks during the vegetative stages, and flowers, fruits, and seeds are the dominant sinks during the reproductive stages (Masclaux-Daubresse et al., 2010; Tegeder et al., 2017). Amino acids, the dominant form of nitrogen, and ureides are transferred into the apoplast and the imported into the seed embryo where they are partition for seed metabolism and storage protein (Tegeder and Rentsch, 2010; Tegeder et al., 2017). Glutamine is the principal nitrogen supply to the cotyledon by contributing 55% of the embryo nitrogen requirement, 20% comes from asparagine, and negligible amount from ureides, allantoin, and allantoic acid (Rainbird et al., 1984). Glycinin and β-conglycinin are the primary storage proteins and account for 55 to 85% of the nitrogen in soybean seed (Murphy and Resurreccion, 1984).

Zhou et al. (2019) studied two soybean cultivars with three different nitrogen treatments: continuous high nitrogen content (CHN), continuous low nitrogen (CLN), and enriched nitrogen supply (ENS) at the R1 stage. ENS at the R1 growth stage increased shoot biomass more than root biomass (Zhou et al., 2019). Soybean demands plentiful nutrition and better canopy for photosynthesis at the R1 stage to increase shoot biomass to benefit reproductive growth (Silva et al., 2011; Zhou et al., 2019). Li et al. (2018) stated that the carbon requirement for organs decreases as plants grow under a prolonged nitrogen limitation, but under high nitrogen condition, plants require more carbon for nitrogen metabolism for plant growth.

Cafaro La Menza et al. (2017) and Cafaro La Menza et al. (2019) developed a protocol to assess nitrogen limitation across a wide range of environments by comparing a control treatment

(zero-nitrogen) where the crop relies on BNF and soil nitrogen supply, and full-nitrogen treatment where the crop is provided with nitrogen fertilizer to match the optimal expected seasonal plant nitrogen demand. Cafaro La Menza et al. (2017) concluded that protein yields were higher in full-nitrogen compared to zero-nitrogen, yield increase is dependent on upon on the yield potential of the environment, and that there is a gap between crop nitrogen requirement and nitrogen supply. To better understand seasonal nitrogen physiological mechanisms (e.g., BNF, aboveground dry matter (ADM) and nitrogen accumulation, leaf area index (LAI), photosynthesis, and nitrogen mobilization), Cafaro La Menza et al. (2019) observed differences in seed yield and seed nitrogen concentration between soybean crops growing under different nitrogen supply and concluded that the amount of nitrogen mobilized was greater in the full-nitrogen treatment vs the zero-nitrogen treatment due to the higher accumulation of nitrogen in the non-seed ADM at the R5 growth stage.

Fabre and Planchon, (2000) studied fourteen F_6 recombinant inbred lines (RIL) to evaluate the sources of nitrogen on yield and seed protein content and study the nitrogen sources across the reproductive growth stages. Fabre and Planchon, (2000) reported a positive associated between seed protein and nitrogen fixation activity from the reproductive stages of R2 – R6 and at R6 plus 10 days. The R5 stage had the highest symbiotic activity for protein production ($r = 0.389$) and for yield, the R6 stage showed a symbiotic activity of $r = 0.549$. During the later reproductive growth stages, nitrogen remobilization and increased nitrogen fixation rates are associated with high seed protein content and can also improve seed yield (Fabre and Planchon, 2000; Leffel et al., 1992).

More than 70 QTL have been identified for nitrogen fixation across all 20 soybean chromosomes (www.soybase.org, accessed 03/01/2021). Bazzer et al. (2020) studied a

population of 196 F_6-derived RIL and identified 13 total quantitative trait loci (QTL) associated with isotope ^{15}N and with a phenotypic variance ranging from 1.63%-14.39%. Isotope ^{15}N best linear unbiased prediction (BLUP) values were used in the QTL mapping because the BLUP values reduces the environmental effect, which increases the accuracy of detection of QTL. Zhou et al. (2017) identified seven QTL for nitrogen uptake, identified six QTL associated with nitrogen use efficiency, eight QTL were detected for biomass yield, and two QTL were identified for grain yield in rice. Dhanapal et al. (2015) identified 17 single nucleotide polymorphism (SNP) associated with nitrogen derived from the atmosphere (NDFA), 19 SNP associated with nitrogen concentration, and 24 SNP associated with carbon/nitrogen (C/N) ratio in soybeans from a genome-wide association study (GWAS) using BLUP values from each environment and the BLUP mean across all environments. The identified SNP were able to indicate 12, 11, and 17 putative loci to be associated with NDFA, plant biomass nitrogen concentration, and the ratio of C/N, respectively.

The objective of this study is to examine the nitrogen mobilization from a mapping population from the cross PI 361103 (contains high shoot N content and low seed N content) x PI 567572B (contains high seed N content and low shoot N content) and to identify QTL for plant biomass whole plant carbon content, whole plant nitrogen content, and seed composition. The secondary objective is to study maternal effects of cytoplasmic inheritance for plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein between population 1 (PI 361103 x PI 567572B) and the population 2 (the reciprocal cross PI 567572B x PI 361103).

Materials and Methods

Population Development

Two mapping populations were developed in 2012 by Dr. James Rusty Smith (USDA-ARS, Stoneville, MS). Population 1 (PI 361103 x PI 567572B) and population 2 (the reciprocal cross PI 567572B x PI 361103) were both developed and advanced with the same procedures. F_2 plants were grown and advanced by single seed descent by Dr. Felix Fritschi (University of Missouri) at Bradford Research Center, Columbia, MO to the F_6 generation. In 2016, F_7 experimental lines were bulked at Bradford Research Center, Columbia, MO. During the 2017 summer field season, Arun Dhanapal (postdoctoral under Dr. Felix Fritschi, University of Missouri) grew $F_{7:8}$ RIL at Bradford Research Center, Columbia, MO and plants were bulked harvested. In 2018, $F_{7:9}$ RIL were grown with three replications at three locations: Rollins Bottom, Columbia, MO, Bradford Research Center, Columbia, MO, and Lee Greenley Memorial Jr. Research Center, Novelty, MO. All experimental plots were planted in single row plots in 9' row with 3' alley and at a depth of 2.5 centimeter with a planting density of 16 seeds per feet.

Genotyping Analysis

DNA was collected from a single replicate in 2018 by collecting young trifoliate leaf tissue from approximately 10 plants per plot at Rollins Bottom, Columbia, MO. A total of 265 DNA samples were isolated using a Maxwell® instrument (Promega Corporation, Madison, WI, USA). and the Maxwell® AS1600 RSC Purefood DNA kit which includes a standard operation procedure for DNA isolation provided by (Promega Corporation, Madison, WI, USA). A total of 240 DNA samples were then submitted to the Soybean Genomics and Improvement Laboratory, USDA-ARS Beltsville, MD, for BARCSoySNP50K BeadChip Illumina Infinium genotyping

array (SoySNP50K) (Song et al., 2013) and an additional 25 samples were submitted for BARCSoySNP6K BeadChip Illumina Infinium genotyping array (SoySNP6K) (Song et al., 2020).

Seed Oil, Seed Protein, and Plant Biomass Analysis

Plants were grown at field four locations, 2017 at Bradford Research Facility Columbia, MO (17B1F), 2018 at Rollins Bottom, Columbia, MO (18ROL), 2018 at Bradford Research Facility, Columbia, MO (18F4B), and 2018 Lee Greenley Memorial Jr. Research Facility, Novelty, MO (18NOV). In this study, we leveraged two different near infrared spectroscopy (NIRS) systems: 1) we analyzed soybean seeds for protein and oil content, and 2) we developed and tested ground whole plant biomass samples for whole plant nitrogen and whole plant carbon content. A total of 262 RIL with three replications from four locations were examined (excluding a small number of missing samples) via NIRS for seed protein and oil content on a dry content basis using a Perten model DA 7250 (Perten Instruments, Hägersten Sweden). NIRS calibrations were developed by Perten Instruments and the University of Minnesota technical staff, using the 2018 NIRS whole seed calibrations (Supplementary Table 3-1).

In the summer of 2017, five soybean shoot samples were collected from 261 single row plots (three replications) from 17B1F for a total of 783 samples in a brown paper bag for population 1. During the summer of 2018, five soybean shoot samples were collected from 262 single row plots (three replications) from 18ROL and 18F4B, and three soybean shoot samples were collected from 262 single row plots (three replications) from 18NOV for a total of 2,358 samples in a brown paper bag for population 1. A total of 783 samples from 261 single row plots (three replications) with five soybean shoots per plot were collected for population 2 in 2018 from Bradford Research Facility, Columbia, MO (18F4B-2). All samples were stored until

relatively dry in greenhouse until moved to storage in an attic at Bradford Research Facility, Columbia, MO. One of two days before taking plant biomass measurements, bags were dried for at least 24 hours using two Precision Quincy quality ovens (Precision Quincy Corporation, Woodstock, IL, USA). All samples were weighed using a LAB Scale PCE-BS 6000 (PCE Americas Inc., Jupiter, FL, USA) and were measured in grams (g). Whole soybean plant tissues were grinded for plant biomass samples using a Viking Electric Hammer Mill Model C-H (Horvick Manufacturing Company, Moorhead, MN, USA) and a Thomas-Wiley Laboratory Mill Model 4 (Arthur H. Thomas Company, Philadelphia, PA, USA) for a coarse grind. Subsamples were then finely grinded using a Cyclone Sample Mill (UDY Corporation, Fort Collins, CO, USA) prior to whole plant carbon and whole plant nitrogen analysis.

NIRS Calibration for Plant Biomass, Carbon, and Nitrogen Content

Ground plant biomass samples from one replicate of two field locations (18NOV replicate 3 and 18ROL replicate 2), as well as all available parental samples from all four 2018 field locations, were selected to develop new NIRS calibrations for whole plant carbon and whole plant nitrogen content. Ground samples were submitted to the University of Missouri Soil Testing and Plant Diagnostic Service Laboratory and total carbon and nitrogen were determined by combustion analysis.

All field collected and ground samples were scanned using a FOSS® 6500 near infrared spectroscopy (NIR-F) reflectance instrument (Foss North America, Eden Prairie, Minnesota). NIR-F spectra for samples were via preprocessed using multiplicative scatter correction (MSC) as described by Geladi et al. (1985) and Savitzky-Golay first derivative (1st Deri) as described by Spielbauer et al. (2009). All spectral preprocessing, partial least squares regression (PLSR) analysis and prediction were carried out using the UnScrambler® version 10.3 (CAMO ASA,

Olav Tryggvason Gt 24, N-7011 Trondheim, Norway). A partial least squares regression method was used to develop calibration models for whole plant carbon (%) and whole plant nitrogen (%) content. In the PLSR analysis, processed spectra data (*X* variable) was used to predict analytical data (*Y* variable). Spectral outliers identified by Unscrambler were removed for the calibrations and final reference ranges for whole plant carbon content (%) were between 38.74-46.00% (n=415 mean 43.17±0.94) and for whole plant nitrogen content (%) were between 2.26-3.68% (n=365, mean 2.90±0.26).

Samples which had both reference values and spectra were randomly assigned to segments (24 for whole plant nitrogen and 25 for whole plant carbon, respectively) so that predicted and measured values could be compared to estimate calibration accuracy through coefficient of correlation (R) as well as standard error of calibration/validation (Table 3-1). The calibration for whole plant nitrogen content was more predictive (R=0.92 for calibration and R=0.9 for validation set) as compared to the calibration for whole plant carbon content (R=0.69 for calibration and R=0.63 for validation set). The standard error of cross validation for whole plant carbon was 0.113 for the calibration and 0.123 for the cross validation. Whole plant nitrogen had a standard error of cross validation of 0.678 for the calibration and 0.734 for the cross validation. Calibrations were then used to predict values for all available spectral samples, and these outputs were used for QTL mapping.

Statistical Analysis of Phenotypic Data

Descriptive statistical analysis was conducted in RStudio version 1.2.1335 (RStudio Team) and the analysis of variance (ANOVA) was conducted using the ‘aov’ function. ANOVA and broad-sense heritability on an entry-mean basis were carried out for plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content of

the three replicated lines from 17B1F, 18ROL, 18F4B, and 18NOV. The ANOVA statistical model is shown below:

$$y_{ijk} = \mu + G_i + G_iE_j + E_j + R_{kj} + e_{ijk}$$

where y_{ijk} represents protein content in the *i*th genotype under the *k*th environment being the *k*th replication within the *j*th environment, μ represents the mean, G_i represents the *i*th genotype, G_iE_j represents the *i*th genotype by *j*th environment interaction, E_j represents the environmental effect, R_k is the *k*th replication within the *j*th environment, and e_{ijk} represents the residual effects (Fehr, 1991; Bernardo, 2020).

Broad-sense heritability on an entry-mean and plot basis were estimated using the formula below:

$$h^2 \text{ (entry mean basis)} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GE}^2}{E} + \frac{\sigma_e^2}{RE}}$$

where h^2 indicated broad-sense heritability, σ_G^2 is the genotypic variance, σ_{GE}^2 is the genotype x environment variance, *E* is the number of environments. σ_e^2 is the error variance, and *R* is the number of replications (Falconer and Mackay, 2009; Fehr, 1991; Bernardo, 2020).

Best linear unbiased prediction were conducted to estimate the phenotypic means of our RIL population in a multiple environment and replication study. The mixed-linear model for BLUP was adopted from (Bernado, 1996) and was conducted using the function 'lmer' in RStudio version 1.2.1335 (RStudio Team) is shown below:

$$y_{ijk} = \mu + G_i + GE_{ij} + E_j + R_{kj} + e_{ijk}$$

where μ is the mean, G_i is the genetic effects of genotypes, GE_{ij} is the genotype and environmental interaction effect, E_j is the environmental effect, R_{kj} is the *k*th replication within the *j*th environment effect, and e_{ik} represents the residual effect. G_i was the fixed effect and GE_{ij}, E_j, and R_{kj} are the random effects (Bernado, 1996).

Population 1 and population 2 produced at the 18ROL location were tested for statistically significant maternal effects from cytoplasmic inheritance differences for all traits: average plant biomass (weight), whole plant carbon content, whole plant nitrogen content, seed protein content, and seed oil content. A t-test was conducted using the function 't.test' in RStudio version 1.2.1335 (RStudio Team) to determine significant means between the two populations for each of the five traits.

SoySNP50K and SoySNP6K Quality Control

Allele calling was conducted using the iScan and GenomeStudio v2.0.5 (Illumina, San Diego, CA, USA). Results were imported into TASSEL version 5.0 (Bradbury et al., 2007) for quality control using modified parameters from Heim et al. (2017) by removing non-segregating and excessively heterozygous (markers greater than 80% heterozygous) and further filtered using the settings of minimum allele frequency at 0.10 and maximum allele frequency at 0.90. Imputation was conducted using the function 'LinkImpute' (Money et al., 2015) method incorporated into TASSAL version 5.0 with the settings = 30 High LD sites, 10 nearest, 10,0000 maximum distance. Parental calling was conducted using the function 'ABH', where A represents parent one allele homozygote (PI361103), B represents parent two allele homozygote (PI567572B), and H represents heterozygous (AB heterozygote), in Tassel version 5.0. Genotypic data were imported into RStudio version 1.2.1335 and the package 'ABHgenotypeR'

(Furuta et al., 2017) was used for error correction using the adjusted of maxHapLength = 5 parameters based of the work of Zhu et al. (2021), resulting in 10,679 total segregating genetic markers.

Genetic Map and Linkage Analysis

The genetic maps and linkage analysis was created in RStudio version 1.2.1335 (RStudio Team) for the population 1 using the 'qtl' package (Broman et al., 2003) for four environments, 17B1F, 18ROL, 18F4B, and 18NOV. After a strict quality control, 10,679 SoySNP50K markers were used to create a genetic map by following a modified pipeline procedure by Dr. Jason Gillman based on the work of Broman, (2011). A total of 4,355 SoySNP50K markers were used for linkage analysis or QTL mapping. QTL mapping was conducted using composite interval mapping using the function 'cim' across 20 chromosomes with the number of marker covariates set at 5, a mapping interval of 10 centimorgan (cM), Expectation-Maximization (EM) algorithm as the mapping method, and error probability of 0.05, for plant biomass content, whole plant carbon content, whole plant nitrogen content, seed protein content, and seed oil content. An empirical logarithm of odds (LOD) thresholds of 4.0 from previous research (Chung et al., 2003; Peiffer et al., 2012; Heim et al., 2017; Patil et al., 2018, Seo et al., 2019) was used as the significant threshold. The effect of each detected QTL were determined by using the function 'sim.geno' with 16 draws and an error probability of 0.001. The QTL effects was used to build an additive QTL model. An additive QTL model for each trait was created, then it was fitted, refined, and then refitted using the functions 'makeqtl', 'refineqtl', and 'fitqtl' (Broman, 2011). Estimated percent variance, and additive effect of each QTL for each trait were extract using the additive QTL model.

Results

Phenotypic Analysis of Seed Oil and Seed Protein

Descriptive phenotypic analysis were conducted using average plant biomass weight, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content for all four environments and for BLUP values (Table 3-2). A mean-separation test between each environment for all the traits concluded that the environments do not group together based on their means. Average plant biomass (g), whole plant carbon content (%), whole plant nitrogen content (%), seed oil content (dry weight basis), and seed protein content (dry weight basis) across four environments and BLUP are shown in Figure 3-1. There was a large variation for biomass between environments and BLUP values ranging from 7.87 g to 54.37g and the coefficient of variation range from 13.49 to 25.54%. Whole plant carbon content ranged from 40.23 to 42.72% and the coefficient of variation ranged from 0.16 to 1.39. The values for whole plant nitrogen content ranges from 1.21 to 3.45% and 4.79 to 11.85 for whole plant nitrogen content and coefficient of variation, respectively. For seed composition, oil content ranged from 13.18 to 18.44% and coefficient of variation ranged from 3.23 to 5.66%, and protein ranged from 39.85 to 41.15% and 2.81 to 3.42% for protein content and coefficient of variation, respectively (Table 3-2).

The distribution of skewness and kurtosis are listed in Table 3-2. The kurtosis value of 18F4B was 4.00, 18NOV for oil was 2.55, and the BLUP for nitrogen was at 2.00. Besides these kurtosis values, the rest are relatively low. In all environments the absolute value of skewness for all traits are less than one. These phenotypic values conform to be continuous distributions and transgressive segregation in the RIL population among all environments. These results indicate

that the RIL population follows a normal distribution model and suggest these traits are controlled by multiple genes. The observed continuous distribution and transgressive segregation demonstrates that alleles from both parents contributed alleles to the RIL population and suggest a random normal population structure for QTL mapping.

Pearson correlation analysis was conducted for all five traits across four environments (as well as BLUP) which are shown in Table 3-3. We observed both positive and negative correlations between all traits across four environments. A consist positive correlation was observed between BLUP and all four environments. These correlations showed significance at $P < 0.05$, $P < 0.01$, and $P < 0.001$ between selected environments. Seed protein and seed oil showed a consist negative correlation across all environments and BLUP. We can conclude that as seed oil content increase, seed protein content will decrease.

An ANOVA and broad-sense heritability on an entry-mean were conducted for the following environments: 17B1F, 18F4B, 18NOV, and 18ROL (Table 3-4). From the ANOVA, environment explained the most variance for all traits. Plant biomass had a large environmental variance of 39,560 and large residual at 38.20. For whole plant carbon and nitrogen content, environmental variance was at 360.61 and 106.26, respectively. As for seed composition, oil content was at 417.74 and protein content was at 48.11. Genotypic variance explained ranged from 0.19 to 59.00 for all traits. Both genotype and environment were both significant at $P<0.001$ for all five traits. Plant biomass, seed protein, and seed oil content were significant at $P<0.001$ for genotype by environment interaction. The entry-mean based heritability for plant biomass was 0.19, whole plant carbon content was 0.33, whole plant nitrogen content was 0.37, seed oil content was 0.80, and seed protein content was 0.83. The results from ANOVA

indicated that environment had the largest impact on the phenotypic values for all five traits and suggest QTL mapping for individual environments.

QTL Identification and Estimated Effects

In this study, a genetic map consisting of 4,355 SNP was used to QTL map plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil, and seed protein content in a RIL mapping population across the four locations 17BIF (Figure 3-2), 18F4B (Figure 3-3), 18NOV (Figure 3-4), 18ROL (Figure 3-5), and BLUP (Figure 3-6). The genetic map span 2,312.5 centimorgan (cM) across 20 chromosomes (Table 3-5). The average length of each chromosome was 115.62 cM, the average spacing between markers was 0.55 cM, and the max spacing between markers ranged from 6 cM to 36.6 cM.

QTL are reported if they were detected in two or more environments including BLUP, have overlapping physical intervals with another trait, and a LOD score above 4.0. The LOD score of 4.0 was set as the significance threshold based on previous studies (Chung et al., 2003; Peiffer et al., 2012; Heim et al., 2017; Patil et al., 2018, Seo et al., 2019). A positive additive effect indicates the allele from the parent PI 361103 and a negative effect indicates the allele from the parent PI 567572B. A total of six QTL for plant biomass were found on Chr. 3, 4, 6, 11, 12, and 15 (Table 3-6). The variance (R^2) explained by these plant biomass QTL ranged from 2.49% (*qB*-4) to 14.26% (*qB*-12) with a LOD values from 4.04 (*qB*-4) to 7.18 (*qB*-15). Out of these QTL, *qB*-3, *qB*-4, *qB*-6, *qB*-12, and *qB*-15 showed a negative additive effect ranging from -0.17 to -0.56. *qB*-3 overlapped with the other detected QTL on Chr. 3 (*qN*-3 and *qPro*-3), *qB*-4 overlapped with *qN*-4, *qB*-6 overlapped with *qOil*-6, *qB*-11 overlapped with *qC*-11 and *qOil*-11, and *qB*-12 overlapped with *qC*-12 and *qPro*-12 (Table 3-7). *qB*-15 was detected in two environments (17B1F and 18ROL) and in BLUP, had the highest LOD score of 7.18, and

explained 3.37% to 3.72% of the variance. *qB*-12 had a LOD score of 4.76 and the highest variance explained at 14.26%.

Two QTL were detected for whole plant carbon content on Chr. 11 (*qC*-11) and Chr. 12 (*qC*-12) with a LOD score of 4.11 and 4.01, respectively (Table 3-6). *qC*-11 overlapped with *qB*-11 and *qOil*-11, and *qC*-12 overlapped with *qPro*-12 and is adjacent to *qB*-12 (Table 3-7). *qC*-11 showed a positive additive effect of 0.31, while *qC*-12 showed a negative additive effect of -0.30. The variance explained was 5.27% and 5.97% for *qC*-11 and *qC*-12, respectively. Three QTL for whole plant nitrogen content were reported on Chr. 3 (*qN*-3), Chr. 4 (*qN*-4), and Chr. 14 (*qN*-14) (Table 3-6). The LOD score ranged from 4.12 to 4.37 and phenotypic variance explained ranged from 1.12% to 11.99%. All three QTL showed a negative effect ranging from -0.15 to -0.35. *qN*-3 overlapped with *qB*-3 and *qPro*-3, *qN*-14 overlapped with *qPro*-14, and *qN*-4 was detected in 18NOV and BLUP and overlapped with *qB*-4 (Table 3-7).

Multiple seed compositional QTL were detected in two or more environments including BLUP and overlapping with other traits. A total of three QTL controlling seed oil content were detected on Chr. 6, 11, and 20, and were named *qOil*-6, *qOil*-11, and *qOil*-20 (Table 3-6). Their phenotypic variation explained 2.35% to 7.22% with LOD scores ranging from 4.48 to 7.39. q*Oil*-6 showed negative additive effects of -0.24 in 17BIF and -0.29 in 18NOV and BLUP. q*Oil*-11 and q*Oil*-20 showed a positive additive effect ranging from 0.26 to 0.39 with only q*Oil*-20 from 18ROL having a negative effect of -0.21. For seed protein content, five QTL were reported on Chr. 3 (*qPro*-3), 12 (*qPro*-12), 13 (*qPro*-13), 14 (*qPro*-14), and 20 (*qPro*-20). *qPro*-3, *qPro*-12, and *qPro*-14 showed a negative additive effect ranging from -0.26 to -0.47 with LOD scores of 4.16 to 9.02 and accounted for 3.95% to 9.47% phenotypic variance explained (Table 3-6). *qPro*-13 and *qPro*-20 LOD scores ranging from 4.32 to 8.54. They showed a positive additive

effect of 0.22 to 0.42 and explained 5.36% to 8.36% of phenotypic variance. *qPro*-3 overlap with *qB*-3 and *qN*-3, *qPro*-12 overlapped with *qC*-12 and *qB*-12, and *qPro*-14 overlapped with *qN*-14 (Table 3-7). Seed oil (*qOil*20) and seed protein (*qPro*-20) on Chr. 20 were detected on identical marker intervals in BLUP and all the environments except for seed oil in 17B1F (Table 3-6). From this result, we suggest that there is an existing protein/oil QTL on Chr. 20.

Maternal Testing for Cytoplasmic Inheritance

Population 1 was a cross between PI 361103 (high shoot nitrogen content and low seed nitrogen content) and PI 567572B (high seed nitrogen content and low shoot nitrogen content). Population 2 was the reciprocal cross (PI 567572B x PI 361103). A t-test between population 1 and population 2 was conducted for plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content (Table 3-8). Both populations were grown in 18ROL and phenotypic values were evaluated the with the same methodology. The mean of whole plant carbon content between the population1 and population 2 was 44.48% and 44.45% with a P-value of 0.54, respectively. For seed composition, oil content was 19.3% for population 1 and 19.74% for population 2 with a P-value 0.1, and protein content was 43.80% and 43.85 for population 1 and population 2 with a P-value of 0.61, respectively. Significant traits were plant biomass and whole nitrogen content at $P < 0.0001$ and $P < 0.001$. The population mean for plant biomass was 16.63 for population 1 and 14.61 for population 2 with a P-value of 6.53E-12. For whole plant nitrogen content, population 1 mean was 2.79% and population 2 mean was 2.74% with a P-value of 0.009. From these results, we suggest that biomass and nitrogen content may be influenced by genes that are inherited maternally through plant organelles.

Discussion

A total of 19 QTL were detected for five traits in this study. We observed an overlap of QTL positions for several traits. QTL for plant biomass, whole plant nitrogen content, and seed protein content overlapped on Chr. 3. QTL for plant biomass and whole plant nitrogen content overlapped on Chr. 4, and plant biomass and seed oil QTL overlapped on Chr. 6. On Chr. 11, QTL for plant biomass, whole plant carbon content, and seed oil content overlapped. QTL for plant biomass, whole plant carbon content, and seed protein content QTL overlapped on Chr. 12. On Chr. 14, the whole plant nitrogen and seed protein content QTL overlapped. Finally, the seed oil and seed protein content QTL on Chr. 20 overlapped. Seed composition is largely affected by carbon partitioning and nitrogen content (Sebolt et al., 2000; Nichols et al., 2006; Kim et al., 2016). A large amount of nitrogen is required for soybean productivity because of high seed protein content (Sinclair and De Wit, 1976; Giller and Cadisch, 1995; Ohyama et al., 2017). These overlapping QTL results may be responsible for seed composition, whole plant carbon, and whole plant nitrogen content in soybean.

In our study, we found three whole plant nitrogen QTL on Chr. 3, Chr. 4, and Chr. 14. Dhanapal et al. (2015) identified one significant SNP on Chr. 3 for nitrogen derived from the atmosphere, and one SNP on Chr.3 and four SNP on Chr. 14 for nitrogen content, which overlapped with our detected QTL. Steketee et al. (2019) reported three significant SNP for nitrogen content on Chr. 3 and two of the SNP overlapped with our QTL on Chr. 3 from 1,720,815 – 34,205,138 bp. The SNP effect ranged from -0.61 to -1.10 and our QTL effect on Chr. 3 was -0.30. These results indicate that our detected whole plant nitrogen QTL may be responsible to total nitrogen content in soybean.

Three QTL for seed oil and five QTL for seed protein were found in our study. The seed protein QTL explained 3.95% to 9.47% of the phenotypic variation, and the seed oil QTL explained 2.35% to 7.22% of the phenotypic variation. Our Chr. 6 seed oil QTL is near the genomic region of cqSeed Oil-016 and cqSeed Oil-016 (Pathan et al., 2013). The Chr. 11 seed oil QTL lies next the reported seed oil QTL from Brummer et al. (1997). Chr. 3 seed protein QTL overlapped in genomic regions with Lee et al. (1996) and is close to cqSeed protein-004 (Pathan et al., 2013). Kabelka et al. (2004) and Zhang et al. (2019) reported a seed protein QTL on Chr. 12 that overlapped with our Chr. 12 seed protein QTL genomic region. Chr. 13 seed protein QTL had an additive effect of 0.29 to 0.42 and overlapped with Seo et al. (2018) novel seed protein QTL. Chr. 14 seed protein QTL overlapped in genomic regions with Diers et al. (1992), Lee et al. (1996) and La, (2018). The seed protein/oil QTL on Chr. 20 was first reported by Diers et al. (1992) and the genetic position was 12 cM and the physical position was from 4,838,960 to 34,233,254. This QTL was later named cqSeed Protein-003 and position at 35.40 to 37.40 cM (Nichols et al., 2006). Our seed protein/oil Chr. 20 QTL was detected in multiple environments and the genetic position was 2.2 – 5.9 cM with a physical interval of 41,867 to 758,718 bp. The seed protein/oil Chr. 20 QTL from our study does not overlap with cqSeed Protein-003 but is instead position on the opposite end of the chromosome. These results indicated that our detected seed oil and seed protein QTL overlapped with many previous studies except on Chr. 20.

The heritability on an entry mean basis for seed protein and seed oil was 0.83 and 0.80, respectively. Compared to Diers et al. (1992) heritability for seed protein and seed oil (0.92 and 0.74, respectively). Heritability of protein and oil content varies depending on the environments and population used (Chung et al., 2003; Hyten et al., 2004).

Maternal trait inheritance through cytoplasmic genetics is when traits are inherited through organelles such as plastids and mitochondria in being transferred directly from the maternal parent to the offspring (Roach and Wulff, 1987). In our study, we conducted a maternal test between the original cross and the reciprocal cross for five traits. Plant biomass and whole plant nitrogen content were significant with a P-value of 6.53E-12 and P-value of 0.009, respectively.

Conclusion

In summary, a total of 19 QTL were detected for plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content from a RIL mapping population between one parent having high shoot nitrogen content and the other parent having high root nitrogen content. Examining whole plant nitrogen and whole plant carbon content in this population can potentially explain the percentage of total seed protein and oil content in soybean seeds. By detecting QTL through linkage analysis, we can uncover the relationship of QTL between these traits. Overlapping QTL may have functions in total carbon and nitrogen content in whole soybean plants which could help explain the partitioning of nitrogen for soybean seed protein content. The reciprocal parental cross and original cross was evaluated for potential maternal inheritance at one location. We concluded that plant biomass and whole plant nitrogen content were significant between the two populations. Further testing needs to be conducted for validation. Whole plant nitrogen and whole plant carbon plays an important role in soybean growth seed composition. Understanding how plants partition nitrogen and carbon, could potentially help plant researchers improve soybean seed composition.

References

Abdelghany AM, Zhang S, Azam M, et al (2019) Natural Variation in Fatty Acid Composition of Diverse World Soybean Germplasms Grown in China. Agronomy 10:24. doi: 10.3390/agronomy10010024

Akond M, Liu S, Boney M, et al (2014) Identification of Quantitative Trait Loci (QTL) Underlying Protein, Oil, and Five Major Fatty Acids' Contents in Soybean. American Journal of Plant Sciences 05:158–167. doi: 10.4236/ajps.2014.51021

Akond M, Liu S, Schoener L, et al (2017) A SNP-Based Genetic Linkage Map of Soybean Using the SoySNP6K Illumina Infinium BeadChip Genotyping Array. Plant Genetics, Genomics, and Biotechnology 1:80–89. doi: 10.5147/pggb.v1i3.154

Alonso-Blanco C, Koornneef M, Stam P (1998) The Use of Recombinant Inbred Lines (RIL) for Genetic Mapping. Arabidopsis Protocols 137–146. doi: 10.1385/0-89603-391-0:137

Andresen JA, Alagarswamy G, Rotz CA, et al (2001) Weather Impacts on Maize, Soybean, and Alfalfa Production in the Great Lakes Region, 1895-1996. Agronomy Journal 93:1059–1070. doi: 10.2134/agronj2001.9351059x

Bazzer SK, Kaler AS, Ray JD, et al (2020) Identification of quantitative trait loci for carbon isotope ratio (δ13C) in a recombinant inbred population of soybean. Theoretical and Applied Genetics 133:2141–2155. doi: 10.1007/s00122-020-03586-0

Beche E, Gillman JD, Song Q, et al (2020) Nested association mapping of important agronomic traits in three interspecific soybean populations. Theoretical and Applied Genetics 133:1039–1054. doi: 10.1007/s00122-019-03529-4

Bernardo R (1994) Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. Crop Science 34:20–25. doi: 10.2135/cropsci1994.0011183x003400010003x

Bernardo R (1996) Best Linear Unbiased Prediction of Maize Single-Cross Performance. Crop Science 36:50–56. doi: 10.2135/cropsci1996.0011183x003600010009x

Bernardo RN (2020) Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN

Boerma HR, Specht JE (2004) Soybeans: improvement, production, and uses. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America

Bohlool BB, Ladha JK, Garrity DP, George T (1992) Biological nitrogen fixation for sustainable agriculture: A perspective. Plant and Soil 141:1–11. doi: 10.1007/bf00011307

Borevitz JO, Nordborg M (2003) The Impact of Genomics on the Study of Natural Variation in Arabidopsis: Figure 1. Plant Physiology 132:718–725. doi: 10.1104/pp.103.023549

Brechenmacher L, Nguyen TH, Zhang N, et al (2015) Identification of Soybean Proteins and Genes Differentially Regulated in Near Isogenic Lines Differing in Resistance to Aphid Infestation. Journal of Proteome Research 14:4137–4146. doi: 10.1021/acs.jproteome.5b00146

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890. doi: 10.1093/bioinformatics/btg112

Broman KW (2004) The Genomes of Recombinant Inbred Lines. Genetics 169:1133–1146. doi: 10.1534/genetics.104.035212

Broman KW (2011) Guide to qtl mapping with r/qtl. Springer-Verlag New York

Brummer EC, Graef GL, Orf J, et al (1997) Mapping QTL for Seed Protein and Oil Content in Eight Soybean Populations. Crop Science 37:370–378. doi: 10.2135/cropsci1997.0011183x003700020011x

Brzostowski LF, Diers BW (2017) Agronomic Evaluation of a High Protein Allele from PI407788A on Chromosome 15 across Two Soybean Backgrounds. Crop Science 57:2972–2978. doi: 10.2135/cropsci2017.02.0083

Brzostowski LF, Pruski TI, Specht JE, Diers BW (2017) Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. Theoretical and Applied Genetics 130:2315–2326. doi: 10.1007/s00122-017-2961-x

Buerkle A, Gompert Z (2012) Population genomics based on low coverage sequencing: how low should we go? Molecular Ecology 22:3028–3035. doi: 10.1111/mec.12105

Cafaro La Menza N, Monzon JP, Specht JE, et al (2019) Nitrogen limitation in high-yield soybean: Seed yield, N accumulation, and N-use efficiency. Field Crops Research 237:74–81. doi: 10.1016/j.fcr.2019.04.009

Cafaro La Menza N, Monzon JP, Specht JE, Grassini P (2017) Is soybean yield limited by nitrogen supply? Field Crops Research 213:204–212. doi: 10.1016/j.fcr.2017.08.009

Carter TE, Nelson RL, Sneller CH, Cui Z (2016) Genetic Diversity in Soybean. Agronomy Monographs 303–416. doi: 10.2134/agronmonogr16.3ed.c8

Chen Q-shan, Zhang Z-chen, Liu C-yan, et al (2007) QTL Analysis of Major Agronomic Traits in Soybean. Agricultural Sciences in China 6:399–405. doi: 10.1016/s1671-2927(07)60062-5

Chung J, Babka HL, Graef GL, 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

Collard BC, Jahufer MZ, Brouwer JB, Pang EC (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169–196. doi: 10.1007/s10681-005-1681-5

Cordeiro CF, Echer FR (2019) Interactive Effects of Nitrogen-Fixing Bacteria Inoculation and Nitrogen Fertilization on Soybean Yield in Unfavorable Edaphoclimatic Environments. Scientific Reports. doi: 10.1038/s41598-019-52131-7

Dei HK (2011) Soybean as a Feed Ingredient for Livestock and Poultry. Recent Trends for Enhancing the Diversity and Quality of Soybean Products. doi: 10.5772/17601

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm.

Dhanapal AP, Ray JD, Singh SK, et al (2015) Genome-Wide Association Analysis of Diverse Soybean Genotypes Reveals Novel Markers for Nitrogen Traits. The Plant Genome. doi: 10.3835/plantgenome2014.11.0086

Diers BW, Keim P, Fehr WR, Shoemaker RC (1992) RFLP analysis of soybean seed protein and oil content. Theoretical and Applied Genetics 83:608–612. doi: 10.1007/bf00226905

Doerge RW, Zeng Z-B, Weir BS (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statistical Science 12:195–219. doi: 10.1214/ss/1030037909

Doyle, Jj & Doyle, Jl. (1986). A Rapid DNA Isolation Procedure from Small Quantities of Fresh Leaf Tissues. Phytochem Bull. 19.

Eskandari M, Cober ER, Rajcan I (2013) Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theoretical and Applied Genetics 126:1677–1687. doi: 10.1007/s00122-013-2083-z

Evangelista JS, Alves RS, Peixoto MA, et al (2021) Soybean productivity, stability, and adaptability through mixed model methodology. Ciência Rural. doi: 10.1590/0103-8478cr20200406

Fabre F, Planchon C (2000) Nitrogen nutrition, yield and protein content in soybean. Plant Science 152:51–58. doi: 10.1016/s0168-9452(99)00221-6

Falconer DS, Mackay TFC (2009) Introduction to quantitative genetics. Pearson, Prentice Hall, Harlow

Fasoula VA, Harris DK, Boerma HR (2004) Validation and Designation of Quantitative Trait Loci for Seed Protein, Seed Oil, and Seed Weight from Two Soybean Populations. Crop Science 44:1218–1225. doi: 10.2135/cropsci2004.1218

FAOSTAT, www.fao.org/faostat/en/#search/soybean. Accessed 3/01/2021.

Fehr WR, Fehr EL, Jessen HJ (1991) Principles of cultivar development. W.R. Fehr, Ames, IA

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proceedings of the National Academy of Sciences 97:4718–4723. doi: 10.1073/pnas.97.9.4718

Fritschi FB, Ray JD, Purcell LC, et al (2013) DIVERSITY AND IMPLICATIONS OF SOYBEAN STEM NITROGEN CONCENTRATION. Journal of Plant Nutrition 36:2111–2131. doi: 10.1080/01904167.2012.748800

Fujikake H (2003) Quick and reversible inhibition of soybean root nodule growth by nitrate involves a decrease in sucrose supply to nodules. Journal of Experimental Botany 54:1379–1388. doi: 10.1093/jxb/erg147

Fujikake H, Yashima H, Sato T, et al (2002) Rapid and reversible nitrate inhibition of nodule growth and N2fixation activity in soybean (Glycine max(L.) Merr.). Soil Science and Plant Nutrition 48:211–217. doi: 10.1080/00380768.2002.10409193

Furuta T, Ashikari M, Jena KK, et al (2017) Adapting Genotyping-by-Sequencing for Rice F2 Populations. G3: Genes|Genomes|Genetics 7:881–893. doi: 10.1534/g3.116.038190

Geladi P, MacDougall D, Martens H (1985) Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. Applied Spectroscopy 39:491–500. doi: 10.1366/0003702854248656

Gelli M, Mitchell SE, Liu K, et al (2016) Mapping QTLs and association of differentially expressed gene transcripts for multiple agronomic traits under different nitrogen levels in sorghum. BMC Plant Biology. doi: 10.1186/s12870-015-0696-x

Giller KE, Cadisch G (1995) Future benefits from biological nitrogen fixation: An ecological approach to agriculture. Management of Biological Nitrogen Fixation for the Development of More Productive and Sustainable Agricultural Systems 255–277. doi: 10.1007/978-94-011-0053-3_13

Guo B, Sleper DA, Arelli PR, et al (2006) Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763. Theoretical and Applied Genetics 112:984–985. doi: 10.1007/s00122-005-0150-9

Haldane JB, Waddington CH (1931) INBREEDING AND LINKAGE. Genetics 16:504–504. doi: 10.1093/genetics/16.5.504a

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324. doi: 10.1038/hdy.1992.131

Hammond EG, Johnson LA, Su C, et al (2005) Soybean Oil. Bailey's Industrial Oil and Fat Products. doi: 10.1002/047167849x.bio041

Harlan JR, de Wet JM, Price EG (1973) Comparative Evolution of Cereals. Evolution 27:311. doi: 10.2307/2406971

Harper JE (1974) Soil and Symbiotic Nitrogen Requirements for Optimum Soybean Production 1. Crop Science 14:255–260. doi: 10.2135/cropsci1974.0011183x001400020026x

Harper JE, Nicholas JC (1978) Nitrogen Metabolism of Soybeans. Plant Physiology 62:662–664. doi: 10.1104/pp.62.4.662

Hartwig EE, Kilen TC (1991) Yield and Composition of Soybean Seed from Parents with Different Protein, Similar Yield. Crop Science 31:290–292. doi: 10.2135/cropsci1991.0011183x003100020011x

Heim CB, Gillman JD (2016) Genotyping-by-Sequencing-Based Investigation of the Genetic Architecture Responsible for a ~Sevenfold Increase in Soybean Seed Stearic Acid. G3: Genes|Genomes|Genetics 7:299–308. doi: 10.1534/g3.116.035741

Huang J, Ma Q, Cai Z, et al (2020) Identification and Mapping of Stable QTLs for Seed Oil and Protein Content in Soybean [Glycine max(L.) Merr.]. Journal of Agricultural and Food Chemistry 68:6448–6460. doi: 10.1021/acs.jafc.0c01271

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

Hymowitz T, Collins FI, Panczner J, Walker WM (1972) Relationship Between the Content of Oil, Protein, and Sugar in Soybean Seed 1. Agronomy Journal 64:613–616. doi: 10.2134/agronj1972.00021962006400050019x

Hyten DL, Pantalone VR, Sams CE, et al (2004) Seed quality QTL in a prominent soybean population. Theoretical and Applied Genetics 109:552–561. doi: 10.1007/s00122-004-1661-5

Jaganathan D, Bohra A, Thudi M, Varshney RK (2020) Fine mapping and gene cloning in the post-NGS era: advances and prospects. Theoretical and Applied Genetics 133:1791–1810. doi: 10.1007/s00122-020-03560-w

Jander G, Norris SR, Rounsley SD, et al (2002) Arabidopsis Map-Based Cloning in the Post-Genome Era. Plant Physiology 129:440–450. doi: 10.1104/pp.003533

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455. doi: 10.1093/genetics/136.4.1447

Johnson HW, Robinson HF, Comstock RE (1955) Genotypic and Phenotypic Correlations in Soybeans and Their Implications in Selection 1. Agronomy Journal 47:477–483. doi: 10.2134/agronj1955.00021962004700100008x

Kabelka EA, Diers BW, Fehr WR, et al (2004) Putative Alleles for Increased Yield from Soybean Plant Introductions. Crop Science 44:784–791. doi: 10.2135/cropsci2004.7840

Kakiuchi J, Kobata T (2008) High Carbon Requirements for Seed Production in Soybeans [Glycine max(L.) Merr.]. Plant Production Science 11:198–202. doi: 10.1626/pps.11.198

Kaur G, Serson W, Orlowski J, et al (2017) Nitrogen Sources and Rates Affect Soybean Seed Composition in Mississippi. Agronomy 7:77. doi: 10.3390/agronomy7040077

Kim M, Schultz S, Nelson RL, Diers BW (2016) Identification and Fine Mapping of a Soybean Seed Protein QTL from PI 407788A on Chromosome 15. Crop Science 56:219–225. doi: 10.2135/cropsci2015.06.0340

La TC, Scaboo A (2018) Characterization of a diverse USDA collection of wild soybean (glycine soja siebold & zucc.) accessions and subsequent mapping for seed composition and agronomic traits in a RIL population. (Doctoral disstertation) Retrieved from https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/66386/research.pdf?sequence=1&isAllowed=y

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199. doi: 10.1093/genetics/121.1.185

Lee J, Hwang Y-S, Kim ST, et al (2017) Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. Food Chemistry 214:248–258. doi: 10.1016/j.foodchem.2016.07.066

Lee SH, Bailey MA, Mian MA, et al (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theoretical and Applied Genetics 93-93:649–657. doi: 10.1007/bf00224058

Leffel RC, Cregan PB, Bolgiano AP, Thibeau DJ (1992) Nitrogen Metabolism of Normal and High-Seed-Protein Soybean. Crop Science 32:747–750. doi: 10.2135/cropsci1992.0011183x003200030034x

Lestari P, Van K, Lee J, et al (2013) Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. Frontiers in Plant Science. doi: 10.3389/fpls.2013.00176

Li H, Ye G, Wang J (2006) A Modified Algorithm for the Improvement of Composite Interval Mapping. Genetics 175:361–374. doi: 10.1534/genetics.106.066811

Li M-W, Muñoz NB, Wong C-F, et al (2016) QTLs Regulating the Contents of Antioxidants, Phenolics, and Flavonoids in Soybean Seeds Share a Common Genomic Region. Frontiers in Plant Science. doi: 10.3389/fpls.2016.00854

Li Y, Yu Z, Jin J, et al (2018) Impact of Elevated CO2 on Seed Quality of Soybean at the Fresh Edible and Mature Stages. Frontiers in Plant Science. doi: 10.3389/fpls.2018.01413

Li, Xu, Yang, Zhao (2019) Dissecting the Genetic Architecture of Seed Protein and Oil Content in Soybean from the Yangtze and Huaihe River Valleys Using Multi-Locus Genome-Wide Association Studies. International Journal of Molecular Sciences 20:3041. doi: 10.3390/ijms20123041

Liu B, Fujita T, Yan Z-H, et al (2007) QTL Mapping of Domestication-related Traits in Soybean (Glycine max). Annals of Botany 100:1027–1038. doi: 10.1093/aob/mcm149

Liu Z, Li H, Fan X, et al (2016) Phenotypic Characterization and Genetic Dissection of Growth Period Traits in Soybean (Glycine max) Using Association Mapping. PLOS ONE. doi: 10.1371/journal.pone.0158602

Lu W, Wen Z, Li H, et al (2012) Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. Theoretical and Applied Genetics 126:425–433. doi: 10.1007/s00122-012-1990-8

Martínez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoretical and Applied Genetics 85:480–488. doi: 10.1007/bf00222330

Masclaux-Daubresse C, Daniel-Vedele F, Dechorgnat J, et al (2010) Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. Annals of Botany 105:1141–1157. doi: 10.1093/aob/mcq028

Masuda, Tadayoshi & Goldsmith, Peter. (2009). World Soybean Production: Area Harvested, Yield, and Long-Term Projections. International Food and Agribusiness Management Review. 12.

Mello Filho OL, Sediyama CS, Moreira MA, et al (2004) Grain yield and seed quality of soybean selected for high protein content. Pesquisa Agropecuária Brasileira 39:445–450. doi: 10.1590/s0100-204x2004000500006

Miao L, Yang S, Zhang K, et al (2019) Natural variation and selection in GmSWEET39 affect soybean seed oil content. New Phytologist 225:1651–1666. doi: 10.1111/nph.16250

Money D, Gardner K, Migicovsky Z, et al (2015) LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3: Genes|Genomes|Genetics 5:2383–2390. doi: 10.1534/g3.115.021667

Morr CV (1981) Nitrogen Conversion Factors for Several Soybean Protein Products. Journal of Food Science 46:1362–1363. doi: 10.1111/j.1365-2621.1981.tb04175.x

Murphy PA, Resurreccion AP (1984) Varietal and environmental differences in soybean glycinin and .beta.-conglycinin content. Journal of Agricultural and Food Chemistry 32:911–915. doi: 10.1021/jf00124a052

Naegle E, Kwanyuen P, Burton J, et al (2008) Seed Nitrogen Mobilization in Soybean: Effects of Seed Nitrogen Content and External Nitrogen Fertility. Journal of Plant Nutrition 31:367–379. doi: 10.1080/01904160801894921

Nascimento D, Polo LR, Lazzari F, et al (2018) Genomic Association between SNP Markers and QTLs for Protein and Oil Content in Grain Weight in Soybean (Glycine max). Journal of Scientific Research and Reports 20:1–13. doi: 10.9734/jsrr/2018/44150

Nichols DM, Glover KD, Carlson SR, et al (2006) Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits. Crop Science 46:834–839. doi: 10.2135/cropsci2005.05-0168

Ohyama T, Minagawa R, Ishikawa S, et al (2013) Soybean Seed Production and Nitrogen Nutrition. A Comprehensive Survey of International Soybean Research - Genetics, Physiology, Agronomy and Nitrogen Relationships. doi: 10.5772/52287

Ohyama T, Tewari K, Ishikawa S, et al (2017) Role of Nitrogen on Growth and Seed Yield of Soybean and a New Fertilization Technique to Promote Nitrogen Fixation and Seed Yield. Soybean - The Basis of Yield, Biomass and Productivity. doi: 10.5772/66743

Pantalone VR, Rebetzke GJ, Burton JW, Wilson RF (1997) Genetic regulation of linolenic acid concentration in wild soybean Glycine soja accessions. Journal of the American Oil Chemists' Society 74:159–163. doi: 10.1007/s11746-997-0162-5

Panter DM, Allen FL (1995) Using Best Linear Unbiased Predictions to Enhance Breeding for Yield in Soybean: I. Choosing Parents. Crop Science 35:397. doi: 10.2135/cropsci1995.0011183x003500020020x

Pathan SM, Vuong T, Clark K, et al (2013) Genetic Mapping and Confirmation of Quantitative Trait Loci for Seed Protein and Oil Contents and Seed Weight in Soybean. Crop Science 53:765–774. doi: 10.2135/cropsci2012.03.0153

Patil G, Chaudhary J, Vuong TD, et al (2017) Development of SNP Genotyping Assays for Seed Composition Traits in Soybean. International Journal of Plant Genomics 2017:1–12. doi: 10.1155/2017/6572969

Patil G, Do T, Vuong TD, et al (2016) Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. Scientific Reports. doi: 10.1038/srep19199

Patil G, Mian R, Vuong T, 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 TD, Kale S, 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

Peiffer, Gregory A., et al. "Identification of Candidate Genes Underlying an Iron Efficiency Quantitative Trait Locus in Soybean." *Plant Physiology*, vol. 158, no. 4, 2012, pp. 1745–1754., doi:10.1104/pp.111.189860.

Pollard DA (2012) Design and Construction of Recombinant Inbred Lines. Methods in Molecular Biology 31–39. doi: 10.1007/978-1-61779-785-9_3

Pratap A, Das A, Kumar S, Gupta S (2021) Current Perspectives on Introgression Breeding in Food Legumes. Frontiers in Plant Science. doi: 10.3389/fpls.2020.589189

Priolli RH, Carvalho CR, Bajay MM, et al (2019) Genome analysis to identify SNPs associated with oil content and fatty acid components in soybean. Euphytica. doi: 10.1007/s10681-019-2378-5

Promega Corporation, Madison, WI, USA. https://www.promega.com/-/media/files/resources/protocols/technical-manuals/101/maxwell-rsc-purefood-gmo-and-authentication-kit-protocol.pdf?la=en. Retrieved from March 17, 2021.

Qi Z-ming, Wu Q, Han X, et al (2011) Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179:499–514. doi: 10.1007/s10681-011-0386-1

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Rainbird RM, Thorne JH, Hardy RW (1984) Role of Amides, Amino Acids, and Ureides in the Nutrition of Developing Soybean Seeds. Plant Physiology 74:329–334. doi: 10.1104/pp.74.2.329

Ray JD, Dhanapal AP, Singh SK, et al (2015) Genome-Wide Association Study of Ureide Concentration in Diverse Maturity Group IV Soybean [Glycine max (L.) Merr.] Accessions. G3: Genes|Genomes|Genetics 5:2391–2403. doi: 10.1534/g3.115.021774

Ray JD, Fritschi FB, Heatherly LG (2006) Large applications of fertilizer N at planting affects seed protein and oil concentration and yield in the Early Soybean Production System. Field Crops Research 99:67–74. doi: 10.1016/j.fcr.2006.03.006

Ray JD, Heatherly LG, Fritschi FB (2006) Influence of Large Amounts of Nitrogen on NoNIRSrigated and Irrigated Soybean. Crop Science 46:52–60. doi: 10.2135/cropsci2005.0043

Rentsch D, Schmidt S, Tegeder M (2007) Transporters for uptake and allocation of organic nitrogen compounds in plants. FEBS Letters 581:2281–2289. doi: 10.1016/j.febslet.2007.04.013

Roach DA, Wulff RD (1987) Maternal Effects in Plants. Annual Review of Ecology and Systematics 18:209–235. doi: 10.1146/annurev.es.18.110187.001233

Salvagiotti F, Cassman KG, Specht JE, et al (2008) Nitrogen uptake, fixation and response to fertilizer N in soybeans: A review. Field Crops Research 108:1–13. doi: 10.1016/j.fcr.2008.03.001

Santachiara G, Borrás L, Salvagiotti F, et al (2017) Relative importance of biological nitrogen fixation and mineral uptake in high yielding soybean cultivars. Plant and Soil 418:191–203. doi: 10.1007/s11104-017-3279-9

Santos MA, Geraldi IO, Garcia AA, et al (2013) Mapping of QTLs associated with biological nitrogen fixation traits in soybean. Hereditas 150:17–25. doi: 10.1111/j.1601-5223.2013.02275.x

Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi: 10.1038/nature08670

Sebolt AM, Shoemaker RC, Diers BW (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

Seo J-H, Kim K-S, Ko J-M, et al (2018) Quantitative trait locus analysis for soybean (Glycine max) seed protein and oil concentrations using selected breeding populations. Plant Breeding 138:95–104. doi: 10.1111/pbr.12659

Sillanpää MJ, Arjas E (1998) Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data. Genetics 148:1373–1388. doi: 10.1093/genetics/148.3.1373

Silva MA, Muniz AS, Mannigel AR, et al (2011) Monitoring and evaluation of need for nitrogen fertilizer topdressing for maize leaf chlorophyll readings and the relationship with grain

yield. Brazilian Archives of Biology and Technology 54:665–674. doi: 10.1590/s1516-89132011000400004

Sinclair TR, de Wit CT (1975) Photosynthate and Nitrogen Requirements for Seed Production by Various Crops. Science 189:565–567. doi: 10.1126/science.189.4202.565

Skoneczka JA, Maroof MA, Shang C, Buss GR (2009) Identification of Candidate Gene Mutation Associated With Low Stachyose Phenotype in Soybean Line PI200508. Crop Science 49:247–255. doi: 10.2135/cropsci2008.07.0403

Song J, Liu Z, Hong H, et al (2016) Identification and Validation of Loci Governing Seed Coat Color by Combining Association Mapping and Bulk Segregation Analysis in Soybean. PLOS ONE. doi: 10.1371/journal.pone.0159064

Song Q, Hyten DL, Jia G, et al (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE. doi: 10.1371/journal.pone.0054985

Song Q, Hyten DL, Jia G, et al (2015) Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3: Genes|Genomes|Genetics 5:1999–2006. doi: 10.1534/g3.115.019000

Song Q, Yan L, Quigley C, et al (2020) Soybean BARCSoySNP6K: An assay for soybean genetics and breeding research. The Plant Journal 104:800–811. doi: 10.1111/tpj.14960

Spain SL, Barrett JC (2015) Strategies for fine-mapping complex traits. Human Molecular Genetics. doi: 10.1093/hmg/ddv260

Spielbauer G, Armstrong P, Baier JW, et al (2009) High-Throughput Near-Infrared Reflectance Spectroscopy for Predicting Quantitative and Qualitative Composition Phenotypes of Individual Maize Kernels. Cereal Chemistry Journal 86:556–564. doi: 10.1094/cchem-86-5-0556

Stein HH, Berger LL, Drackley JK, et al (2008) Nutritional Properties and Feeding Values of Soybeans and Their Coproducts. Soybeans 613–660. doi: 10.1016/b978-1-893997-64-6.50021-4

Steketee CJ, Sinclair TR, Riar MK, et al (2019) Unraveling the genetic architecture for carbon and nitrogen related traits and leaf hydraulic conductance in soybean using genome-wide association analyses. BMC Genomics. doi: 10.1186/s12864-019-6170-7

Streeter J, Wong PP (1988) Inhibition of legume nodule formation and N2fixation by nitrate. Critical Reviews in Plant Sciences 7:1–23. doi: 10.1080/07352688809382257

UnScrambler® version 10.3 (CAMO ASA, Olav Tryggvason Gt 24, N-7011 Trondheim, Norway

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of Quantitative Trait Loci for Protein and Lipid Contents in Soybean Seeds Using Recombinant Inbred Lines. Breeding Science 53:133–140. doi: 10.1270/jsbbs.53.133

Tamagno S, Balboa GR, Assefa Y, et al (2017) Nutrient partitioning and stoichiometry in soybean: A synthesis-analysis. Field Crops Research 200:18–27. doi: 10.1016/j.fcr.2016.09.019

Tamagno S, Sadras VO, Haegele JW, et al (2018) Interplay between nitrogen fertilizer and biological nitrogen fixation in soybean: implications on seed yield and biomass allocation. Scientific Reports. doi: 10.1038/s41598-018-35672-1

Tegeder M, Masclaux-Daubresse C (2017) Source and sink mechanisms of nitrogen transport and use. New Phytologist 217:35–53. doi: 10.1111/nph.14876

Tegeder M, Rentsch D (2010) Uptake and Partitioning of Amino Acids and Peptides. Molecular Plant 3:997–1011. doi: 10.1093/mp/ssq047

Truong Q, Koch K, Yoon JM, et al (2013) Influence of carbon to nitrogen ratios on soybean somatic embryo (cv. Jack) growth and composition. Journal of Experimental Botany 64:2985–2995. doi: 10.1093/jxb/ert138

Uga Y, Sugimoto K, Ogawa S, et al (2013) Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. Nature Genetics 45:1097–1102. doi: 10.1038/ng.2725

Van K, McHale L (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 18:1180. doi: 10.3390/ijms18061180

Wang J, Chen P, Wang D, et al (2015) Identification and mapping of stable QTL for protein content in soybean seeds. Molecular Breeding. doi: 10.1007/s11032-015-0285-6

Wang P-wu, Di Q, Liu X-Y (2020) Genome-Wide association Study Identifies Candidate Genes Related to Oleic acid content of Soybean Seed. doi: 10.21203/rs.3.rs-17853/v1

Warrington CV, Abdel-Haleem H, Hyten DL, 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

Watanabe S, Xia Z, Hideshima R, et al (2011) A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals that theGIGANTEAGene Is Involved in Soybean Maturity and Flowering. Genetics 188:395–407. doi: 10.1534/genetics.110.125062

Whittaker JC, Thompson R, Visscher PM (1996) On the mapping of QTL by regression of phenotype on marker-type. Heredity 77:23–32. doi: 10.1038/hdy.1996.104

Wilson, R.F. 2004. Seed composition. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, Production, and Uses. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.: 621-677

Wood CW, Torbert HA, Weaver DB (1993) Nitrogen Fertilizer Effects on Soybean Growth, Yield, and Seed Composition. Journal of Production Agriculture 6:354–360. doi: 10.2134/jpa1993.0354

Xu G, Fan X, Miller AJ (2012) Plant Nitrogen Assimilation and Use Efficiency. Annual Review of Plant Biology 63:153–182. doi: 10.1146/annurev-arplant-042811-105532

Yang Q, Yang Y, Xu R, et al (2019) Genetic Analysis and Mapping of QTLs for Soybean Biological Nitrogen Fixation Traits Under Varied Field Conditions. Frontiers in Plant Science. doi: 10.3389/fpls.2019.00075

Ye H, Song L, Chen H, et al (2018) A major natural genetic variation associated with root system architecture and plasticity improves waterlogging tolerance and yield in soybean. Plant, Cell & Environment. doi: 10.1111/pce.13190

Young ND, Zamir D, Ganal MW, Tanksley SD (1988) Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the tm-2a gene in tomato. Genetics 120:579–585. doi: 10.1093/genetics/120.2.579

Yu X, Yuan F, Fu X, Zhu D (2016) Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. Food Chemistry 196:776–782. doi: 10.1016/j.foodchem.2015.09.092

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468. doi: 10.1093/genetics/136.4.1457

Zhang J, Wang X, Lu Y, et al (2018) Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. Molecular Plant 11:460–472. doi: 10.1016/j.molp.2017.12.016

Zhang T, Wu T, Wang L, et al (2019) A Combined Linkage and GWAS Analysis Identifies QTLs Linked to Soybean Seed Protein and Oil Content. International Journal of Molecular Sciences 20:5915. doi: 10.3390/ijms20235915

Zhang W-K, Wang Y-J, Luo G-Z, et al (2004) QTL mapping of ten agronomic traits on the soybean (Glycine max L. Merr.) genetic map and their association with EST markers. Theoretical and Applied Genetics 108:1131–1139. doi: 10.1007/s00122-003-1527-2

Zhang YH, Liu MF, He JB, et al (2015) Marker-assisted breeding for transgressive seed protein content in soybean [Glycine max (L.) Merr.]. Theoretical and Applied Genetics 128:1061–1072. doi: 10.1007/s00122-015-2490-4

Zhou H, Yao X, Zhao Q, et al (2019) Rapid Effect of Nitrogen Supply for Soybean at the Beginning Flowering Stage on Biomass and Sucrose Metabolism. Scientific Reports. doi: 10.1038/s41598-019-52043-6

Zhou Y, Tao Y, Tang D, et al (2017) Identification of QTL Associated with Nitrogen Uptake and Nitrogen Use Efficiency Using High Throughput Genotyped CSSLs in Rice (Oryza sativa L.). Frontiers in Plant Science. doi: 10.3389/fpls.2017.01166

Zhu X, Leiser WL, Hahn V, Würschum T (2021) Identification of seed protein and oil related QTL in 944 RILs from a diallel of early-maturing European soybean. The Crop Journal 9:238–247. doi: 10.1016/j.cj.2020.06.006

Table 3-1. NIRS calibration and cross validation for estimating whole plant nitrogen, and whole carbon content.

Parameter	Nitrogen	Carbon
	Calibration	
n[a]	365	415
Spectra[b]	400-2490	400-2490
PLS factors[c]	13	13
SECV[d]	0.113	0.678
RMSECV[e]	0.113	0.677
R[f]	0.92	0.69
R^{2g}	0.84	0.48
Measured range	2.26-3.68%	38.74-46.00%
mean±SD	2.90±0.26%	43.17±0.94%
	Cross validation	
n[a]	365	415
Spectra[b]	400-2490	400-2490
PLS factors[c]	13	13
SECV[d]	0.123	0.734
RMSECV[e]	0.123	0.733
R[f]	0.90	0.63
R^{2g}	0.81	0.40
SD[h]	1.9	1.3
Predicted range	0.11-3.86%	38.44-48.72%
mean±SD	2.66±0.48%	43.75±1.04%
# CV segments	24	25

[a]Number of samples; [b]Spectrum measured in nanometer (nm); [c]Partial Least Squares factors; [d]Standard error of cross validation; [e]Root-Mean-Square error of cross validation; [f]Correlation coefficient; [g]Coefficient of determination of cross validation; [h]Standard Deviation

Table 3-2. Descriptive statistical analysis and mean separation groupings for average plant biomass (g), whole plant carbon content (%), whole plant nitrogen content (%), seed oil content (dry weight basis), and seed protein content (dry weight basis) across four environments and BLUP values.

traits	environment	min[b]	max[b]	mean[b]	SD[c]	CV (%)[d]	skewness	kurtosis	groups[e]
biomass	17B1F	7.87	25.49	14.07	2.64	18.76	0.62	0.96	c
	18F4B	8.49	28.59	16.73	3.67	21.96	0.73	0.42	b
	18NOV	13.99	54.37	27.01	6.90	25.54	0.79	1.06	a
	18ROL	5.51	18.30	9.82	2.23	22.70	0.66	0.71	d
	BLUP[a]	11.21	23.37	16.89	2.28	13.49	0.38	0.16	
carbon	17B1F	42.28	45.52	44.04	0.58	1.32	-0.29	-0.07	b
	18F4B	42.20	46.20	44.48	0.51	0.16	-0.29	1.69	a
	18NOV	42.17	44.34	43.32	0.35	0.80	-0.03	0.42	c
	18ROL	40.23	44.22	42.80	0.60	1.39	-0.42	0.99	d
	BLUP[a]	42.72	44.43	43.66	0.29	0.67	-0.17	0.34	
nitrogen	17B1F	1.21	2.75	2.14	0.25	11.85	-0.44	0.60	d
	18F4B	1.59	3.41	2.79	0.24	8.44	-0.98	4.00	b
	18NOV	1.82	3.26	2.72	0.22	8.03	-0.80	2.38	c
	18ROL	2.38	3.45	3.03	0.18	5.96	-0.36	0.44	a
	BLUP[a]	2.12	3.04	2.67	0.13	4.79	-0.31	2.00	
oil	17B1F	18.44	21.54	19.88	0.64	3.23	-0.12	-0.44	a
	18F4B	17.59	21.99	19.81	0.81	4.08	0.03	-0.10	a
	18NOV	13.18	21.59	18.34	1.04	5.66	-0.60	2.55	b
	18ROL	17.59	21.81	19.86	0.82	4.12	0.02	-0.34	a
	BLUP[a]	17.15	21.09	19.48	0.66	3.40	-0.17	0.06	
protein	17B1F	40.97	47.29	43.94	1.23	2.81	0.14	-0.20	b
	18F4B	40.39	48.43	43.80	1.46	3.34	0.12	-0.16	c
	18NOV	39.85	49.25	44.41	1.52	3.42	0.05	0.21	a
	18ROL	40.55	47.33	43.99	1.38	3.15	0.07	-0.28	b
	BLUP[a]	41.15	17.40	44.03	1.14	2.58	0.05	0.15	

[a]Best Linear Unbiased Prediction; [b]Minimum, Maximum, and Mean; [c]Standard Deviation; [d]Coefficient of variation in percentage; [e]Mean separation groups

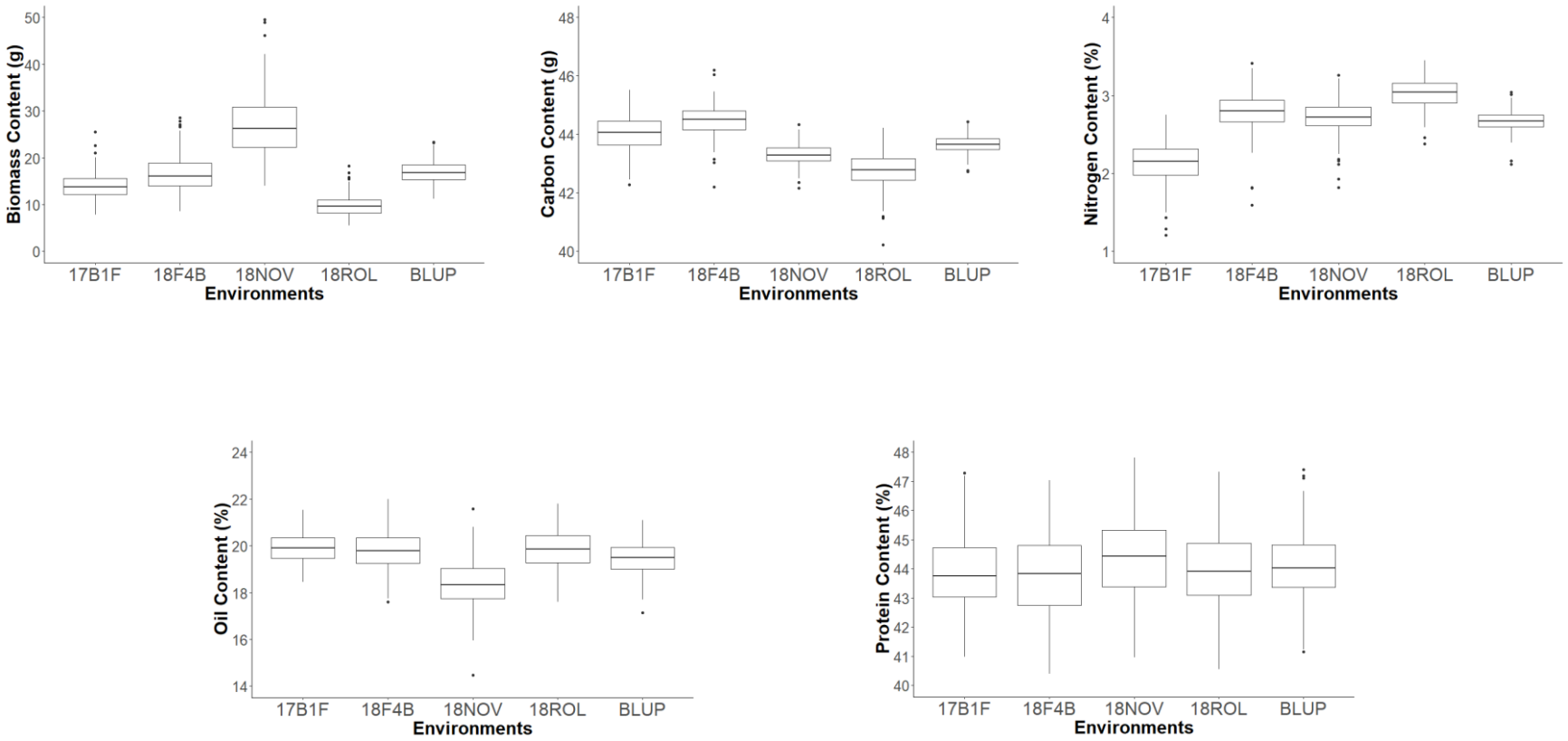

Figure 3-1. Average plant biomass (g), whole plant carbon content (%), whole plant nitrogen content (%), seed oil content (dry weight basis), and seed protein content (dry weight basis) across four environments and BLUP.

Table 3-3. Pearson correlation of plant biomass (B), whole plant carbon content (C), whole plant nitrogen content (N), seed oil content (Oil), and seed protein content (Pro) across four environments and BLUP.

		17B1F					18F4B					18NOV					18ROL					BLUP				
env.	trait	B	C	N	Oil	Pro	B	C	N	Oil	Pro	B	C	N	Oil	Pro	B	C	N	Oil	Pro	B	C	N	Oil	Pro
17B1F	B	1																								
	C	0.182**	1																							
	N	-0.007	-0.272***	1																						
	Oil	0.088	-0.108*	-0.037	1																					
	Pro	-0.003	0.148*	0.087	-0.541***	1																				
18F4B	B	0.065	0.044	-0.003	-0.089	0.130*	1																			
	C	0.008	0.083	-0.054	-0.002	-0.081	-0.228***	1																		
	N	-0.069	-0.038	0.069	-0.058	0.088	0.058	-0.298***	1																	
	Oil	0.061	-0.049	-0.052	0.545***	-0.267***	-0.034	-0.039	-0.089	1																
	Pro	-0.009	0.046	0.120*	-0.435***	0.522***	0.055	-0.127*	0.178**	-0.674***	1															
18NOV	B	0.025	0.001	0.017	-0.029	0.020	0.102	-0.094	0.109*	-0.018	0.044	1														
	C	-0.055	0.068	-0.022	0.057	0.019	0.070	0.079	-0.173**	0.007	-0.028	0.002	1													
	N	0.053	-0.087	0.264	-0.078	0.097	0.054	-0.012	0.146*	-0.018	0.074	0.074	-0.139*	1												
	Oil	0.008	-0.006	-0.042	0.503***	-0.344***	-0.119*	-0.032	-0.043	0.482***	-0.404***	0.037	0.058	-0.045	1											
	Pro	-0.025	0.022	0.081	-0.405***	0.533***	0.156*	-0.076	0.145*	-0.382***	0.554***	0.062	-0.013	0.036	-0.785***	1										
18ROL	B	-0.047	-0.039	0.062	-0.086	0.134	0.125*	-0.015	0.147*	-0.025	0.035	-0.010	-0.053	-0.046	-0.031	0.086	1									
	C	0.156*	0.137*	-0.108*	-0.016	0.056	-0.037	0.039	-0.144*	-0.032	0.100	0.088	0.054	0.008	0.032	0.000	-0.376***	1								
	N	0.024	-0.042	0.064	-0.072	0.063	0.025	-0.181**	0.148*	-0.033	0.117*	-0.067	-0.026	-0.065	-0.049	0.075	0.169**	0.173**	1							
	Oil	0.055	-0.059	-0.083	0.514***	-0.305***	-0.031	-0.067	-0.101	0.616***	-0.500***	-0.055	0.078	-0.040	0.502***	-0.420***	0.046	0.010	0.086	1						
	Pro	-0.020	0.058	0.137	-0.376***	0.438***	0.016	-0.093	0.199**	-0.481***	0.657***	0.073	-0.023	0.056	-0.405***	0.556***	0.020	0.053	0.012	-0.764	1					
BLUP	B	0.329***	0.060	0.027	-0.067	0.103	0.490***	-0.148*	0.125*	-0.025	0.062	0.822***	-0.009	0.077	-0.036	0.121*	0.274***	0.003	0.011	-0.035	0.067	1				
	C	0.003	0.070	-0.032	-0.146	0.028	0.095	0.046	-0.126*	-0.091	0.006	-0.104	0.084	0.011	-0.159	0.118*	0.012	0.005	-0.004	-0.025	-0.045	-0.046	1			
	N	-0.017	-0.225**	0.721**	-0.099	0.129	0.043	-0.205**	0.487***	-0.094	0.205**	0.037	-0.127*	0.627***	-0.074	0.132*	0.120*	-0.056	0.386***	-0.079	0.186**	0.070	-0.060	1		
	Oil	0.056	-0.059	-0.070	0.758***	-0.433***	-0.083	-0.047	-0.089	0.816***	-0.623***	-0.016	0.065	-0.051	0.809***	-0.647***	-0.025	-0.002	-0.020	0.815***	-0.632***	-0.047	-0.126*	-0.106*	1	
	Pro	-0.016	0.077	0.134	-0.533***	0.741***	0.109*	-0.117*	0.193**	-0.563***	0.851***	0.060	-0.019	0.075	-0.606***	0.824***	0.080	0.065	0.089	-0.615***	0.821***	0.1059*	0.029	0.204**	-0.725***	1

*Indicates significant at the 0.1 level ($P < 0.1$)

**Indicates significant at the 0.05 level ($P < 0.05$)

***Indicates significant at the 0.05 level ($P < 0.01$)

Table 3-4. The analysis of variance and heritability on an entry-mean basis for plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content.

Source of Variance		Mean Square					Pr(>F)				
	DF	Biomass	Carbon	Nitrogen	Oil	Protein	Biomass	Carbon	Nitrogen	Oil	Protein
Genotype (G)	259	59.00	0.92	0.19	4.96	14.53	3.97E-07*	1.34E-08*	1.28E-06*	<2.22E-16*	<2.22E-16*
Environment (E)	3	39560.00	360.61	106.26	417.74	48.11	<2.22E-16*	<2.22E-16*	<2.22E-16*	<2.22E-16*	<2.22E-16*
Genotype x Environment (GxE)	776	47.90	0.62	0.12	0.97	2.51	6.24E-05*	0.08	0.87	6.53E-06*	2.34E-08*
Replications in Environment	8	442.50	21.73	0.67	1.13	3.48	3.38E-16*	<2.22E-16*	1.12E-06*	0.13	0.04
Residual	1850	38.20	0.57	0.12	0.75	1.82					
h^2 (entry mean basis)		0.19	0.33	0.37	0.80	0.83					

*Indicates significant at the 0.001 level ($P < 0.001$)

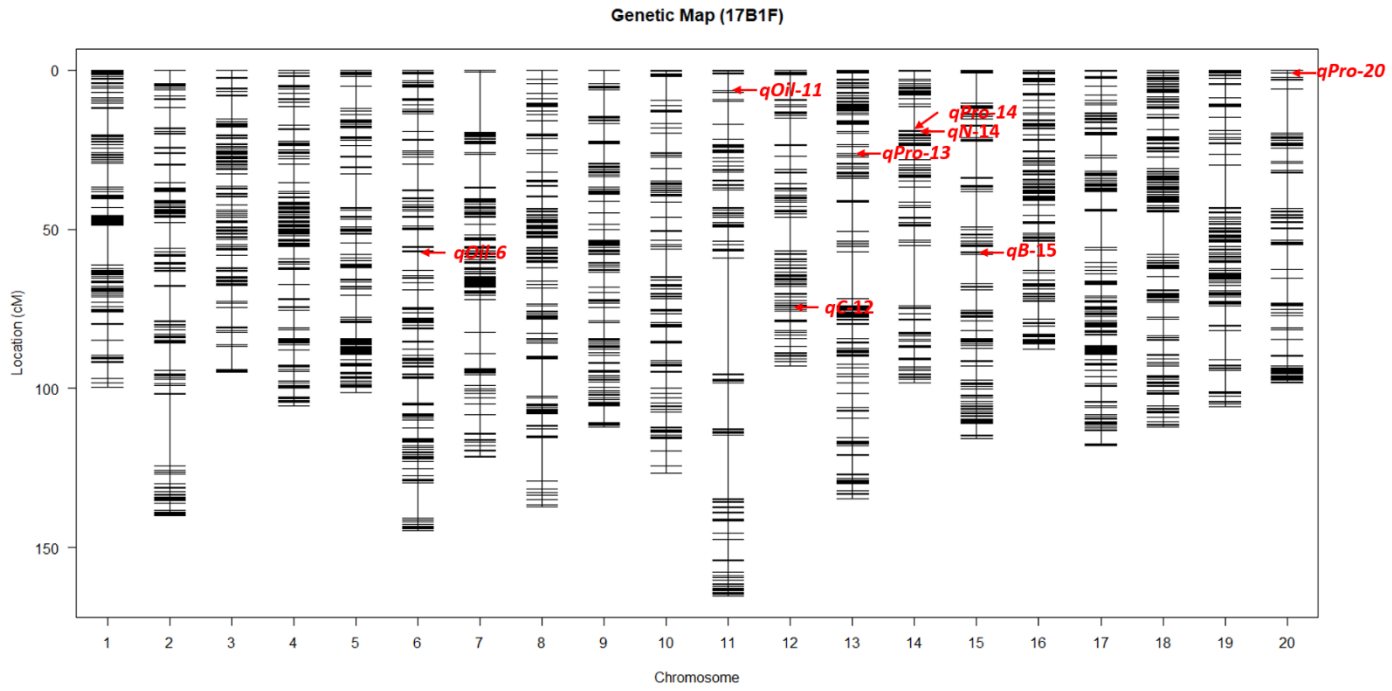

Figure 3-2. The genetic map of 17B1F consisting of 4,355 SNP markers across 20 chromosomes and displaying eight QTL.

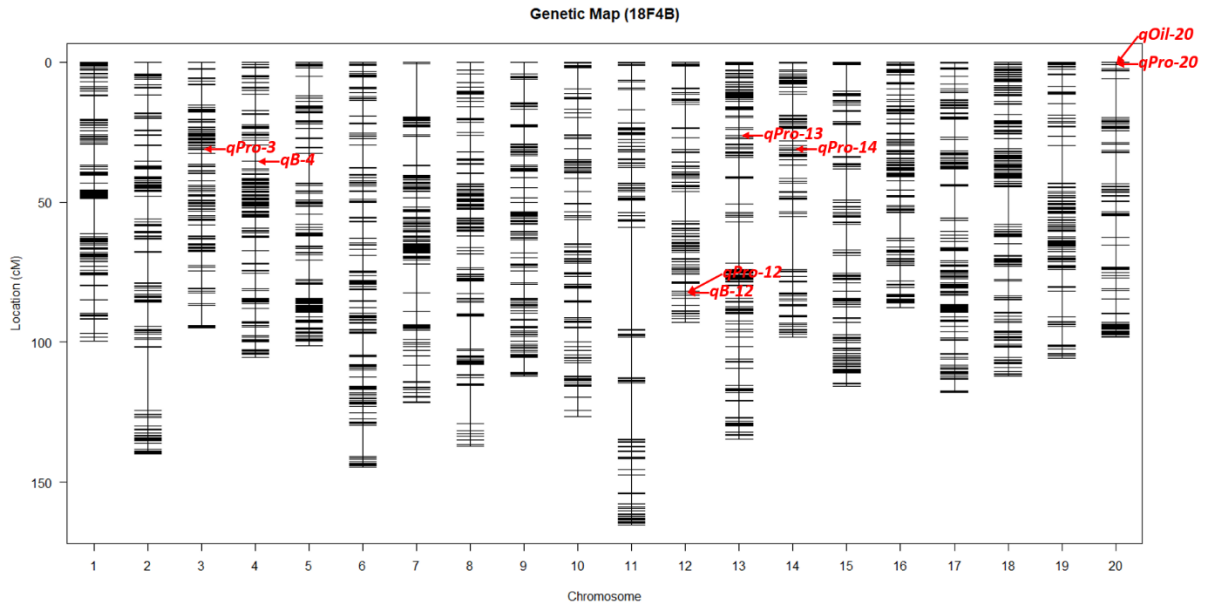

Figure 3-3. The genetic map of 18F4B consisting of 4,355 SNP markers across 20 chromosomes and displaying eight QTL.

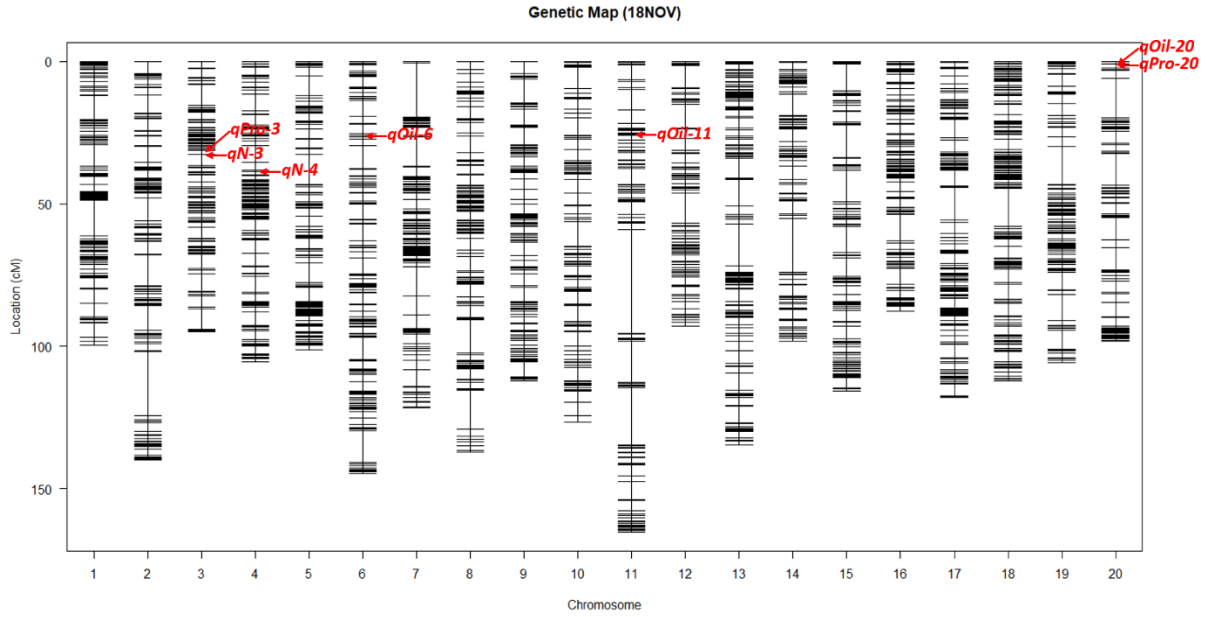

Figure 3-4. The genetic map of 18NOV consisting of 4,355 SNP markers across 20 chromosomes and displaying seven QTL.

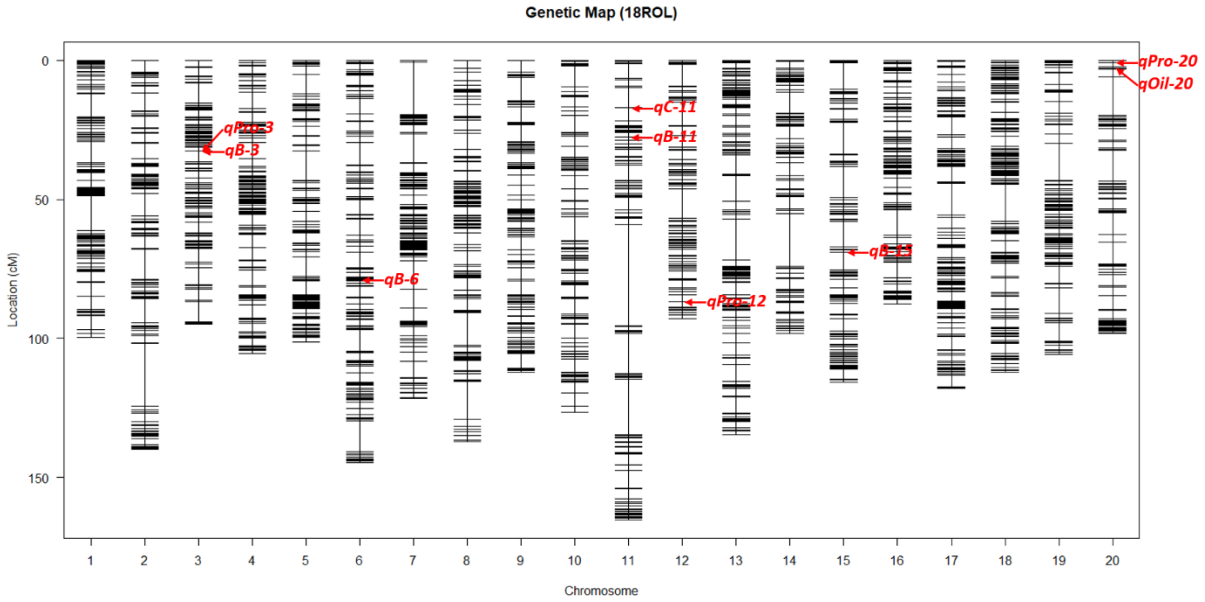

Figure 3-5. The genetic map of 18ROL consisting of 4,355 SNP markers across 20 chromosomes and displaying nine QTL.

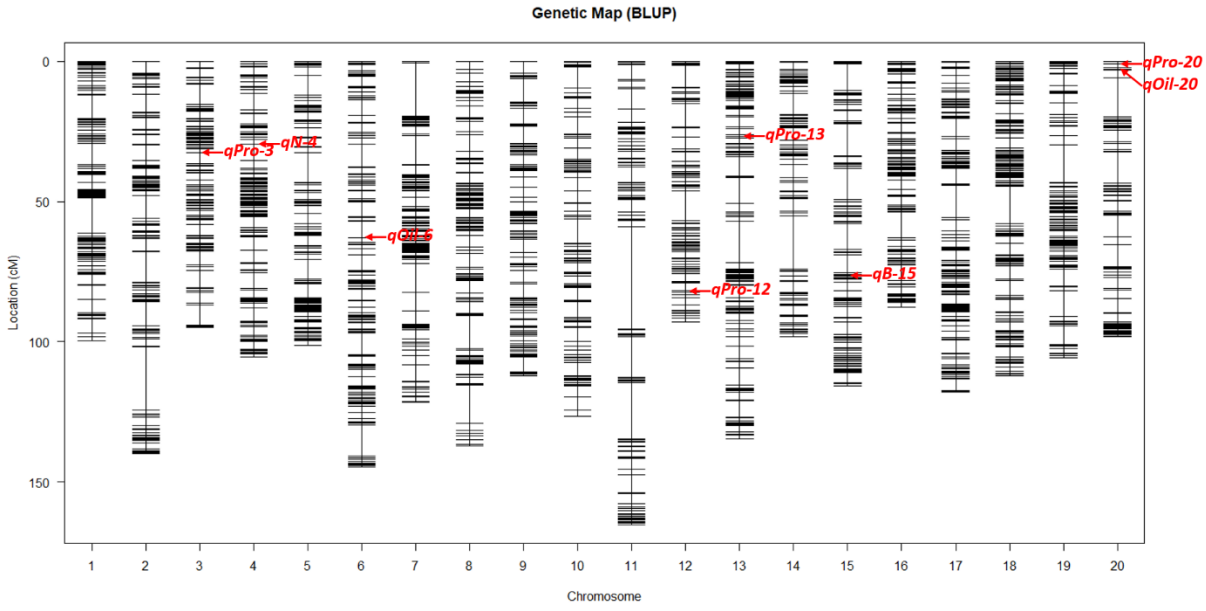

Figure 3-6. The genetic map of BLUP consisting of 4,355 SNP markers across 20 chromosomes and displaying eight QTL.

Table 3-5. SNP marker distribution across 20 chromosomes with number of markers per chromosome, length (cM), average spacing between markers, and max spacing between markers.

	Chr.	Number of Markers	Length	Average Spacing	Max Spacing
	1	197	99.6	0.5	12.5
	2	216	140.1	0.7	22.5
	3	228	94.8	0.4	7.4
	4	209	105.5	0.5	6
	5	207	101.4	0.5	10.6
	6	234	144.7	0.6	10.9
	7	246	121.7	0.5	18.8
	8	238	137.1	0.6	13.7
	9	181	112.3	0.6	8.3
	10	202	126.6	0.6	8.8
	11	194	165.2	0.9	36.6
	12	168	93	0.6	10.6
	13	276	134.8	0.5	14.5
	14	181	98.2	0.5	18.8
	15	239	115.7	0.5	11.4
	16	236	87.7	0.4	9.2
	17	245	117.9	0.5	11.5
	18	316	112	0.4	13.5
	19	195	105.8	0.5	13.5
	20	147	98.3	0.7	13.9
overall		4355	2312.5	0.5	36.6

Table 3-6. QTL mapping of plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content from population 1 in four environments and BLUP.

trait	QTL name[a]	Chr.[b]	marker interval[c]	peak	position (cM)	LOD[d]	A[e]	R^2 (%)[f]	env.[e]	Ref.[h]
biomass	*qB-3*	3	Gm03_3617955-Gm03_44923331	Gm03_7848774	29.8	4.71	-0.30	5.97	18ROL	
	qB-4	4	Gm04_989803-Gm04_52263519	Gm04_7054273	35.4	4.04	-0.17	2.49	18F4B	
	qB-6	6	Gm06_84193-Gm06_51169689	Gm06_13658325	78.4	5.80	-0.20	2.77	18ROL	
	qB-11	11	Gm11_1720815-Gm11_34234546	Gm11_4418072	25.1	4.61	0.27	4.20	18ROL	
	qB-12	12	Gm12_38552678-Gm12_38552678	Gm12_38552678	86.9	4.76	-0.56	14.26	18F4B	
	qB-15	15	Gm15_127608-Gm15_51241958	Gm15_12743349	67.2	7.18	-0.22	3.37	17B1F	
			Gm15_122135-Gm15_50103701	Gm15_15420057	77.0	4.73	-0.27	3.55	BLUP	
			Gm15_2683646-Gm15_51565010	Gm15_12971833	68.0	5.51	-0.20	3.72	18ROL	
carbon	*qC-11*	11	Gm11_1720815-Gm11_34205138	Gm11_3959854	21.9	4.11	0.31	5.27	18ROL	
	qC-12	12	Gm12_36308899-Gm12_38798071	Gm12_37092076	78.8	4.01	-0.30	5.97	17B1F	
nitrogen	*qN-3*	3	Gm03_1991155-Gm03_17767575	Gm_03_7731589	29.5	4.37	-0.35	5.57	18NOV	Dhanapal et al., 2015; Steketee et al., 2019
	qN-4	4	Gm04_989803-Gm04_51725446	Gm_04_6016181	27.0	4.13	-0.29	11.99	BLUP	Dhanapal et al., 2015
		4	Gm04_989803-Gm04_52263519	Gm_04_7289247	38.1	4.32	-0.15	1.12	18NOV	Dhanapal et al., 2015
	qN-14	14	Gm14_1241606-Gm14_48932740	Gm14_4514257	19.0	4.12	-0.22	3.17	17B1F	Dhanapal et al., 2015
oil	*qOil-6*	6	Gm06_1785765-Gm06_48466050	Gm06_3392346	21.8	5.09	-0.29	4.07	18NOV	Pathan et al., 2013
			Gm06_2118361-Gm06_48725006	Gm06_11974449	65.2	4.87	-0.29	4.01	BLUP	Pathan et al., 2013
			Gm06_3104699-Gm06_51169689	Gm06_10277748	55.5	5.77	-0.24	4.08	17B1F	Pathan et al., 2013
	qOil-11	11	Gm11_17705-Gm11_33838162	Gm11_2139177	9.7	4.61	0.26	4.68	17B1F	Brummer et al., 1997
			Gm11_1530847-Gm11_32742820	Gm11_3959854	21.8	5.06	0.34	3.21	18NOV	
	qOil-20	20	Gm20_41867-Gm20_758718	Gm20_476829	2.2	6.51	0.38	7.16	BLUP	
			Gm20_41867-Gm20_758718	Gm20_476829	2.2	7.39	0.39	7.22	18F4B	
			Gm20_208950-Gm20_758718	Gm20_476829	2.2	4.48	0.38	3.79	18NOV	
			Gm20_208950-Gm20_758718	Gm20_476829	2.2	5.64	-0.21	2.35	18ROL	
protein	*qPro-3*	3	Gm03_1438203-Gm03_17767575	Gm03_7731589	29.5	7.41	-0.34	5.03	18NOV	Lee et al., 1996
			Gm03_5032050-Gm03_44923331	Gm03_8079197	30.3	4.16	-0.27	3.95	18ROL	Lee et al., 1996
			Gm03_5764913-Gm03_17767575	Gm03_8079197	30.3	5.84	-0.40	7.54	BLUP	Lee et al., 1996
			Gm03_5764913-Gm03_17767575	Gm03_8079197	30.3	5.86	-0.39	7.03	18F4B	Lee et al., 1996
	qPro-12	12	Gm12_38040866-Gm12_38552678	Gm12_38552678	86.9	9.02	-0.47	9.47	BLUP	Kabelka et al., 2004; Zhang et al., 2019
			Gm12_38040866-Gm12_38552678	Gm12_38552678	86.9	8.42	-0.47	9.47	18F4B	Kabelka et al., 2004; Zhang et al., 2019
			Gm12_38040866-Gm12_38782388	Gm12_38552678	86.9	5.92	-0.42	9.15	18ROL	Kabelka et al., 2004; Zhang et al., 2019
	qPro-13	13	Gm13_10359814-Gm13_18955770	Gm13_16925180	24.1	4.81	0.29	5.36	17B1F	Seo et al., 2018
			Gm13_16841783-Gm13_18552568	Gm13_16925180	24.1	6.39	0.42	8.36	BLUP	Seo et al., 2018
			Gm13_16841783-Gm13_18552568	Gm13_16925180	24.1	7.06	0.41	8.13	18F4B	Seo et al., 2018
	qPro-14	14	Gm14_2177440-Gm14_48932740	Gm14_4514257	19.0	4.22	-0.26	4.26	17B1F	Diers et al., 1992
			Gm14_3081409-Gm14_9039674	Gm14_7246631	33.5	5.06	-0.32	5.30	18F4B	Diers et al., 1992; La, 2018
	qPro-20	20	Gm20_41867-Gm20_758718	Gm20_476829	2.2	4.91	0.35	8.24	17B1F	
			Gm20_41867-Gm20_758718	Gm20_541850	2.9	8.54	0.39	7.17	BLUP	
			Gm20_41867-Gm20_758718	Gm20_608620	3.1	6.32	0.22	7.24	18F4B	
			Gm20_208950-Gm20_758718	Gm20_476829	2.2	5.20	0.41	7.43	18NOV	
			Gm20_208950-Gm20_758718	Gm20_758718	5.9	4.32	0.33	5.64	18ROL	

[a]Designated QTL name; [b]Chromosome; [c]Marker interval and physical interval; [d]Logarithm of the odds; [e]Additive effect; [f]Coefficient of determination; [e]Environment; [h]Reference of previous reported QTL

Table 3-7. Table of overlapping QTL.

trait	QTL name[a]	Chr.[b]	marker interval[c]	env.[d]
protein	*qPro-3*	3	Gm03_1438203-Gm03_17767575	18NOV
nitrogen	*qN-3*	3	Gm03_1991155-Gm03_17767575	18NOV
biomass	*qB-3*	3	Gm03_3617955-Gm03_44923331	18ROL
protein	*qPro-3*	3	Gm03_5032050-Gm03_44923331	18ROL
protein	*qPro-3*	3	Gm03_5764913-Gm03_17767575	BLUP
protein	*qPro-3*	3	Gm03_5764913-Gm03_17767575	18F4B
nitrogen	*qN-4*	4	Gm04_989803-Gm04_51725446	BLUP
biomass	*qB-4*	4	Gm04_989803-Gm04_52263519	18F4B
nitrogen	*qN-4*	4	Gm04_989803-Gm04_52263519	18NOV
oil	*qOil-6*	6	Gm06_1785765-Gm06_48466050	18NOV
oil	*qOil-6*	6	Gm06_2118361-Gm06_48725006	BLUP
oil	*qOil-6*	6	Gm06_3104699-Gm06_51169689	17B1F
biomass	*qB-6*	6	Gm06_84193-Gm06_51169689	18ROL
oil	*qOil-11*	11	Gm11_1530847-Gm11_32742820	18NOV
carbon	*qC-11*	11	Gm11_1720815-Gm11_34205138	18ROL
biomass	*qB-11*	11	Gm11_1720815-Gm11_34234546	18ROL
oil	*qOil-11*	11	Gm11_17705-Gm11_33838162	17B1F
carbon	*qC-12*	12	Gm12_36308899-Gm12_38798071	17B1F
protein	*qPro-12*	12	Gm12_38040866-Gm12_38552678	BLUP
protein	*qPro-12*	12	Gm12_38040866-Gm12_38552678	18F4B
protein	*qPro-12*	12	Gm12_38040866-Gm12_38782388	18ROL
biomass	*qB-12*	12	Gm12_38552678-Gm12_38552678	18F4B

protein	*qPro-13*	13	Gm13_10359814-Gm13_18955770	17B1F
protein	*qPro-13*	13	Gm13_16841783-Gm13_18552568	BLUP
protein	*qPro-13*	13	Gm13_16841783-Gm13_18552568	18F4B
nitrogen	*qN-14*	14	Gm14_1241606-Gm14_48932740	17B1F
protein	*qPro-14*	14	Gm14_2177440-Gm14_48932740	17B1F
protein	*qPro-14*	14	Gm14_3081409-Gm14_9039674	18F4B
biomass	*qB-15*	15	Gm15_122135-Gm15_50103701	BLUP
biomass	*qB-15*	15	Gm15_127608-Gm15_51241958	17B1F
biomass	*qB-15*	15	Gm15_2683646-Gm15_51565010	18ROL
oil	*qOil-20*	20	Gm20_41867-Gm20_758718	BLUP
oil	*qOil-20*	20	Gm20_41867-Gm20_758718	18F4B
protein	*qPro-20*	20	Gm20_41867-Gm20_758718	17B1F
protein	*qPro-20*	20	Gm20_41867-Gm20_758718	BLUP
protein	*qPro-20*	20	Gm20_41867-Gm20_758718	18F4B
oil	*qOil-20*	20	Gm20_208950-Gm20_758718	18NOV
oil	*qOil-20*	20	Gm20_208950-Gm20_758718	18ROL
protein	*qPro-20*	20	Gm20_208950-Gm20_758718	18NOV
protein	*qPro-20*	20	Gm20_208950-Gm20_758718	18ROL

[a]Designated QTL name; [b]Chromosome; [c]Marker interval and physical interval; [d]Environment

Table 3-8. Mean separation test for maternal inheritance from a T-Test between population 1 (PI 361103 x PI 567572B) and population 2 (the reciprocal cross PI 567572B x PI 361103) for plant biomass, whole plant carbon content, whole plant nitrogen content, seed oil content, and seed protein content.

	biomass		carbon		nitrogen		oil		protein	
populations	mean	*P*-value	mean	*P*-value	mean	*P*-value	mean	*P*-value	mean	*P*-value
Population 1	16.63	6.53E12**	44.48	0.54	2.79	0.009*	19.83	0.1	43.8	0.61
Population 2	14.61		44.45		2.74		19.74		43.85	

*Indicates significant at the 0.01 level ($P < 0.01$)
**Indicates significant at the 0.001 level ($P < 0.001$)

Supplementary Table 3-1. Whole soybean seed NIRS calibrations for 2018.

Parameter	2018 Oil	2018 Protein
Calibration	Honigs	Honigs
Type	Regression	Regression
Moisture basis	Dry basis	Dry basis
Parameter Unit	%	%
SECV[a]	0.84	0.86
R^2CV[b]	0.93	0.94
Min[c]	9.92	33.44
Max[c]	26.47	53.55
n[d]	3537	3790
Date[e]	2018	2018

[a]Standard error of cross validation; [b]Coefficient of determination of cross validation; [c]Minimum and maximum value; [d]Number of samples; [e]Last updated date